TREBALL DE FI DE GRAU

Grau en Matemàtiques

Títol: Anàlisi del rendiment dels estudiants del Grau de Matemàtiques de la UPC, amb perspectiva de gènere

Autor: Marc Estévez Inglada

Director: Professora Marta Pérez-Casany

Departament: Estadística i Investigació Operativa

Convocatòria: Juny 2020

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

A la Marta Pérez, directora d'aquest treball de fi de grau, per la seva constant dedicació i el suport que m'ha proporcionat durant la realització d'aquest projecte. Gràcies per oferir-me aquesta proposta de treball tan interessant i que m'ha servit per continuar aprenent aspectes del món de la modelització.

A l'Institut de Ciències de l'Educació (ICE) per la BECA rebuda per a la realització d'aquest projecte.

Als serveis informàtics de la FME, especialment al Jordi Aguilar, pel seu treball en proporcionar unes dades tan clares.

A la meva família i amics pel seu recolzament durant la realització d'aquest treball.

Abstract

It is known that in the last years, the number of women enrolled in the Mathematics Degree at UPC has decreased considerably. Based on that, and on the importance to show to society that both men and women are equally able to pursue the same fields of academia, profession, sports etc, the objective of this study is to compare the performance of the students enrolled in the Bachelor's in Mathematics at UPC from a gender point of view.

The examined dataset consists of 725 students that were enrolled in the Bachelor's in Mathematics at UPC during the school years 2009/10 to 2019/20. The response variables considered in this study are the bachelor's access grade, the results of the level test applied to the students when they are admitted, the mean grade of the initial phase, the number of semesters required to pass the initial phase, number of subjects passed on the first attempt and the dropout probability. Gender, if the student is or not CFIS, the bachelor's access grade and the lowest access grade, the results of the level test and the year of enrolment have been used as possible explanatory variables.

The methodologies that have been applied are Linear Models, Generalized Linear Models with a Binomial response and the analysis of contingency tables.

Keywords

Gender perspective, STEAM, Linear Models, Generalized Linear Models

Índex

1. Introducció

En els darrers anys el nombre d'estudiants dones al Grau de Matemàtiques de la UPC ha anat baixant de forma considerable. Anteriorment al període analitzat, els gèneres estaven força equilibrats, al voltant del 50% dels estudiants eren dones i l'altre 50% eren homes. Durant el curs 2009/2010, que és el curs més antic inclòs en l'estudi, les dones representaven un 38% dels estudiants. El darrer any escolar 2019/2020, només un 24% dels estudiants que es van matricular al grau eren dones, 14 punts percentuals menys respecte del curs 2009/2010.

Dins del període que engloba l'estudi, s'ha pogut apreciar una disminució considerable del nombre de dones al Grau. Aquesta disminució es desconeix si és conseqüència de que les dones troben altres graus més interessants i que, per tant, la seva motivació ha anat canviant al llarg del temps, o bé per altra banda, s'han anat desanimant amb aquest tipus d'estudis.

Aquest treball s'ha portat a terme finançat per una BECA de l'Institut de Ciències de l'Educació (ICE) de la UPC de 5 mesos de duració. L'objectiu principal del mateix és estudiar quina influència té el gènere en diferents variables de rendiment en els estudis de matemàtiques de la UPC. Les metodologies estadístiques emprades en les anàlisis que s'han vist a l'assignatura d'Estadística del Grau han estat les comparacions de mitjanes i variàncies de dues mostres independents, l'anàlisi de taules de contingència i els models lineals. Com a metodologia nova hi figuren els models lineals generalitzats. Les anàlisis s'han fet amb el *software* estadístic R (https://www.r-project.org).

La memòria s'organitza de la forma següent: als Secció 2 i 3 s'expliquen respectivament les metodologies dels models lineals i lineals generalitzats. Al Secció 4 s'explica la base de dades de la qual disposem. El Secció 5 està destinat a l'estudi dels estudiants al moment d'entrar al Grau, per tal de veure si hi ha diferències significatives entre els coneixements dels homes i els de les dones. Al Secció 6 s'analitzen quines variables influencien la variable la variable nota *mitjana de la fase inicial*. Al Secció 7 es realitza el mateix estudi que al Secció 6, però aquesta vegada tenint en compte també la nota de la *prova de nivell*. Als Secció 8 i 9 s'analitzen les taules de contingència per veure si hi ha diferència entre els homes i les dones en el nombre de quadrimestres que necessiten per superar la fase inicial i en el nombre d'assignatures obligatòries aprovades a la primera en el Grau respectivament. Al Secció 10 s'estudia, mitjançant models lineals generalitzats, la probabilitat de que un estudiant no abandoni els estudis, donades les seves particulars característiques. Finalment, el Secció 11 conté les principals conclusions a les que s'han arribat. L'Apèndix A conté el codi en R de les anàlisis realitzades.

2. Metodologia: Models Lineals

En aquesta secció parlarem dels Models Lineals, explicarem com es defineixen, com s'ajusten i com es comprova si el model és bo per a unes dades en particular. La informació d'aquesta secció s'ha tret bàsicament del llibre [1] i dels apunts de l'assignatura d'Estadística del Grau de Matemàtiques de la FME.

2.1 Introducció

L'objectiu dels models lineals consisteix en explicar el comportament d'una variable aleatòria Y, anomenada variable dependent, en funció d'unes certes variables independents $X_1, ..., X_{p-1}$. Aquests, s'expressen de la següent forma:

$$Y = X\beta + e,$$

que matricialment equival a:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1(p-1)} \\ 1 & x_{21} & x_{22} & \cdots & x_{2(p-1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{n(p-1)} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

on $Y_i \sim N(\mu_i, \sigma^2)$, Y_i és independent de Y_j $\forall i \neq j$. A més a més, X és una matriu de n files per p columnes, on n és el nombre de mostres que tenim i p és el numero de paràmetres β que utilitzem en el model, anomenant a β_0 com a intercept. Finalment, tenim que els errors e_i segueixen una distribució $N(0, \sigma^2)$. Encara que en alguns casos alguna de les variables explicatives pugui ser aleatòria, sovint es considera determinista i cadascuna de les files de la matriu X s'interpreta com les condicions experimentals sota les quals s'ha observat la variable Y.

Podem trobar una definició alternativa per als models lineals utilitzant l'esperança de la variable dependent, que és la següent:

$$\mu = \mathbb{E}[Y] = X\beta, \tag{1}$$

on $\mu = \mathbb{E}[Y]$, $Y_i \sim N(\mu_i, \sigma^2)$, Y_i és independent de Y_j $\forall i \neq j$ i la matriu X està igualment definida que en la definició anterior. Aquesta darrera formulació serà útil per a definir, més endavant, els models lineals generalitzats.

Un cop recollides les dades, tindrem un vector $y^t = (y_1, y_2, ..., y_n)$ que serà una realització del vector Y. Una vegada observat el vector Y en les diferents condicions experimentals, ens queden p paràmetres β_i i el paràmetre σ^2, als quals serà necessari donar un valor de manera que ajusti el model de la millor manera possible les dades obtingudes. La següent secció està dedicada a l'estimació dels paràmetres del model.

2.2 Estimació dels paràmetres

Primer de tot, volem estimar el valor del paràmetre β. Sigui $y^t = (y_1, y_2, ..., y_n)$ una realització de Y i $\hat{\beta}$ una estimació de β. Tenim dos maneres d'estimar el valor de β. La primera consisteix en utilitzar el criteri

de *mínims quadrats* que consisteix en minimitzar la funció

$$S(\beta) = \|y - \hat{y}\|_2^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \sum_{j=0}^{p-1} x_{ij}\beta_j)^2$$

Observis que aplicar mínims quadrats equival a minimitzar la discrepància entre el vector d'observacions y i el de prediccions $\hat{y} = X\hat{\beta}$. Calculant $\frac{\partial S(\beta)}{\partial \beta_i}$ per $i = 1, ..., n$ i resolent el sistema $\frac{\partial S(\beta)}{\partial \beta_i} = 0$ per $i = 1, ..., n$ trobem que l'estimació del paràmetre β ve donada per $\hat{\beta} = (X^tX)^{-1}X^ty$.

Una segona forma de trobar una estimació de β consisteix en trobar l'estimació màxim versemblant. Això vol dir que calcularem la funció de versemblança i agafarem com $\hat{\beta}$ el valor del paràmetre que faci més versemblant obtindre les dades que hem obtingut. A diferència de mínims quadrats que és un mètode purament algebraic, màxima versemblança sí que requereix assumir una distribució de probabilitat pel vector resposta.

Calculem primer la funció versemblança de la distribució normal, que atès que $\mu_i = \sum_{j=0}^{p-1} x_{ij}\beta_j$ és igual a:

$$L(\beta; y) = (\sqrt{2\pi}\sigma)^{-n} exp(-\sum_{i=1}^{n} \frac{(y_i - \sum_{j=0}^{p-1} x_{ij}\beta_j)^2}{2\sigma^2})$$

Calculem ara la funció de log-versemblança, que és igual al logaritme de la versemblança i és igual a:

$$l(\beta; y) = -nlog(\sqrt{2\pi}\sigma) - \sum_{i=1}^{n} \frac{(y_i - \sum_{j=0}^{p-1} x_{ij}\beta_j)^2}{2\sigma^2}$$

Atès que el màxim d'una funció és igual al màxim del seu logaritme, busquem el vector β que maximitzi $l(\beta; y)$. A tal efecte trobem el vector denominat *score* $U = (U_1, U_2, \cdots, U_{p-1})$, on

$$U_j = \frac{\partial l}{\partial \beta_j} = \frac{1}{\sigma^2}(X^t(Y - X\beta))_j \ \forall j$$

Finalment, per trobar l'estimador imposem que $U_j = 0 \ \forall j \iff X^tY = X^tX\beta$, la qual cosa dóna com a resultat

$$\hat{\beta} = (X^tX)^{-1}X^tY$$

Observis que el $\hat{\beta}$ trobat per a màxima versemblança és el mateix que havíem trobat fent mínims quadrats. Aquesta és una condició que només es compleix quan s'assumeix que la variable resposta és normal.

Per tal d'estimar la variància de la variable resposta, que és la mateixa que la de la variable error, definirem què s'entén per error residual. En aquest punt és important observar que així com s'assumeix que $\mu_i = \mathbb{E}[Y_i]$ canvia al canviar les condicions experimentals, la variància s'assumeix constant. Aquesta hipòtesi d'igualtat de variàncies es coneix amb el nom d'hipòtesi d'*homocedasticitat* i caldrà comprovar-la més endavant a través de l'anàlisi dels residus.

Definició 2.1. L'error residual és $\hat{e} = Y - X\hat{\beta}$, que és el vector que conté les discrepàncies entre els valors observats i els predits pel model.

Per trobar un estimador de la variància utilitzarem el mètode dels moments.

Sigui $y^t = (y_1, y_2, \cdots, y_n)$ una realització de Y. Primer de tot, sabem que en el cas de tenir una mostra d'una $N(\mu, \sigma^2)$, $\sum_{i=1}^{n} \frac{(y_i-\mu)^2}{\sigma^2} \sim \mathcal{X}_n^2$. Al estimar l'esperança per la mitjana aritmètica es perd un grau de

llibertat i es compleix que $\sum_{i=1}^{n} \frac{(y_i-\bar{y})^2}{\sigma^2} \sim \mathcal{X}^2_{n-1}$. En el nostre cas, i assumint que μ_i és diferent per a cada observació i que s'estimen a través de p paràmetres, tindrem que

$$\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{\sigma^2} \sim \mathcal{X}^2_{n-p}$$

Aplicant ara el mètode dels moments, sabent que $\mathbb{E}[\mathcal{X}^2_{n-p}] = n - p$, tenim que

$$\mathbb{E}[\sum_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{\sigma^2}] = n - p \iff \mathbb{E}[S^2] = \sigma^2,$$

sent

$$S^2 = \frac{1}{n-p} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

l'estimador del paràmetre σ^2. Observis que S^2 equival a fer una mitjana dels errors al quadrat, d'aquí que també s'anomeni *error quadràtic mig* (EQM). Per tant doncs, l'estimador de la variància que es pren és l'error quadràtic mig que obtindrem un cop ajustat el model a les nostres dades. Important esmentar que l'estimador màxim versemblant de la variància dels errors és lleugerament diferent a l'obtingut pel mètode dels moments, sobretot si el nombre d'observacions és petit. El del mètode dels moments però és sense biaix, és a dir, es compleix que $\mathbb{E}[S^2] = \sigma^2$, i per això és el que es pren habitualment.

Un cop es tenen els estimadors puntuals dels paràmetres del model, cal realitzar la inferència respecte a aquests paràmetres. Concretament, serà interessant saber si els β_i són zero o no, perquè en cas de no ser estadísticament diferents de zero el model es podrà simplificar.

2.3 Inferència respecte als paràmetres del model

Ara que ja tenim els paràmetres estimats, volem saber si les variables associades afecten realment al model o, altrament, les podem eliminar d'aquest. Per a saber-ho, realitzarem el següent test d'hipòtesis:

$$\begin{cases} H_0 : \beta_i = 0 \\ H_1 : \beta_i \neq 0 \end{cases} \quad i = 1, ..., p-1 \tag{2}$$

Per poder realitzar aquest test d'hipòtesis, requerim prèviament de la següent proposició.

Proposició 2.2. *El vector $\hat{\beta}$, que és un vector aleatori, té la següent distribució:*

$$\hat{\beta}|X \sim N(\beta, \sigma^2(X^tX)^{-1})$$

Demostració. Primer de tot, sabem que $\hat{\beta}$ és un vector normal, ja que és una combinació lineal de variables aleatòries normals.
Calculant ara l'esperança, tenim

$$\mathbb{E}[\hat{\beta}|X] = (X^tX)^{-1}X^t\mathbb{E}[Y|X] = (X^tX)^{-1}X^tX\beta = \beta,$$

per tant, $\hat{\beta}$ és un estimador sense biaix de β. Calculem ara la variància:

$$\mathbb{E}[(\hat{\beta}|X - \beta)(\hat{\beta}|X - \beta)^t] = (X^tX)^{-1}X^t\mathbb{E}[(Y - X\beta)(Y - X\beta)^t]X(X^tX) = \sigma^2(X^tX)^{-1}$$

Per tant, tindrem que $\hat{\beta}|X$ seguirà una distribució normal amb els paràmetres calculats. □

Donada la Proposició 2.2, tenim que cada component del vector de paràmetres β satisfà $\hat{\beta}_i|X \sim N(\beta_i, \sigma^2[(X^tX)^{-1}]_{ii})$. Utilitzant aquesta propietat al test d'hipòtesis (2), tenim que rebutjarem la hipòtesi nul·la del test a nivell α quan

$$|\frac{\hat{\beta}_i}{S\sqrt{[(X^tX)^{-1}]_{ii}}}| \geq t_{1-\alpha/2,n-p},$$

on $t_{1-\alpha/2,n-p}$ és el punt que acumula una probabilitat igual a $1-\alpha/2$ en una distribució t-Student amb $n-p$ graus de llibertat.

Altrament, no rebutjarem la hipòtesi nul·la i això voldrà dir que la covariable X_i no té una influència significativa en Y i, per tant, la podrem eliminar del model. A continuació trobem una figura[1] que mostra la zona de rebuig del test a nivell α en el cas de tenir 15 graus de llibertat i per $\alpha = 0.05$, on $t_{0.975,15} = 2.131$.

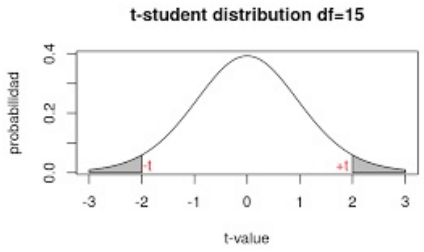


Figura 1: Gràfica de la zona de rebuig de la distribució t-Student amb 15 graus de llibertat

Un altre test que és útil per veure si el model serveix per explicar una part significativa de la variabilitat que hi ha a la variable resposta consisteix en comparar-lo amb el model nul. Aquest test és conegut com test d'*Omnibus* i es defineix de la següent forma:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0 \\ H_1 : \exists i | \beta_i \neq 0 \end{cases} \tag{3}$$

Per realitzar aquest test, necessitem primer unes definicions prèvies.

Definició 2.3. En models lineals intervenen tres sumes de quadrats:

1. Suma total de quadrats = TSS = $\sum_{i=1}^{n}(Y_i - \bar{Y})^2$. Aquesta suma equival a mesurar la variabilitat de les Y_i sense tenir en compte les covariables.
2. Suma residual de quadrats = RSS = $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$. Aquesta suma ja ha sortit abans a l'hora d'estimar σ^2, i correspon a una mesura de la discrepància existent entre valors observats i predits.
3. Suma de quadrats de la regressió = RegSS = $\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$. Aquesta suma de quadrats mesura quina part de la variabilitat de les Y_i és explicada pel nostre model. Observis que $\bar{Y}$ correspon als valors predits sota el model nul.

Es compleix que

$$TSS = RegSS + RSS$$

[1] Gràfica extreta de *https://rpubs.com/Joaquin_AR/218467*

Aquestes sumes de quadrats compleixen que, si H_0 és certa, $\frac{TSS}{\sigma^2} \sim \mathcal{X}^2_{n-1}$ i $\frac{RegSS}{\sigma^2} \sim \mathcal{X}^2_{p-1}$. També es compleix que $\frac{RSS}{\sigma^2} \sim \mathcal{X}^2_{n-p}$ independentment de si estem sota la hipòtesi nul·la o l'alternativa. A més a més, es té que TSS té n-1 graus de llibertat, RSS en té n-p i RegSS en té p-1.

Per tant, per realitzar el test (3), tenim que si H_0 és cert, aleshores

$$F_0 = \frac{RegSS/(p-1)}{RSS/(n-p)} \sim F_{p-1,n-p}$$

i, per tant, rebutjarem el test a nivell α quan $F_0 \geq F_{1-\alpha,p-1,n-p}$, on $F_{1-\alpha,p-1,n-p}$ és el punt de la distribució F de Fisher amb p - 1 i n - p graus de llibertat que acumula una probabilitat per sota igual a $1-\alpha$. A continuació es mostra en la Figura 2[2] com funciona el test a nivell α per la distribució F de Fisher.

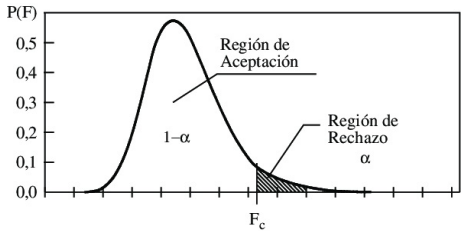


Figura 2: Gràfica de la zona de rebuig de la distribució F de Fisher

Observis que el que interessa per a tenir un bon model és que RegSS sigui gran i RSS sigui petit. Per tant, quan RegSS sigui gran, també ho serà F_0 i tindrà sentit rebutjar la hipòtesi nul·la que diu que el nostre model no explica, perquè sí que explicarà.

2.4 Bondat d'ajust del model

Els mètodes de bondat d'ajust ens permeten veure si el model ajusta adequadament les observacions o, pel contrari, no és prou precís.

Pel que fa a models lineals, el *coeficient de determinació* és una mesura de bondat d'ajust que es defineix amb la següent fórmula:

$$R^2 = \frac{RegSS}{TSS} = 1 - \frac{RSS}{TSS}$$

El coeficient de determinació es pot interpretar com la proporció de variabilitat en la resposta explicada pel nostre model i com més s'apropa a 1, millor s'ajusta a les observacions. També es pot interpretar com 1 menys la variabilitat en la variable resposta no explicada per les covariables.

És evident que com més paràmetres tinguem en el model, millor aproximades estaran les dades i, per tant, més s'aproparà el valor de R^2 a 1. Per això, per comparar models amb diferent nombre de paràmetres i penalitzar aquells que n'utilitzen un nombre més gran, podem calcular el R^2 *ajustat*, que es defineix amb la següent fórmula:

$$R^2_{adj} = 1 - \frac{RSS/(n-p)}{TSS/(n-1)} = 1 - (1-R^2)\frac{n-1}{n-p}$$

[2] Gràfica extreta de *https://www.researchgate.net/figure/Figura-2-Funcion-de-Densidad-de-Probabilidad-para-una-distribucion-F-de-Fisher-de-cola_fig1_261710841*

D'aquesta manera podem trobar en cas que existeixi un model que sigui prou proper a les dades i que, a més a més, utilitzi un nombre relativament petit de paràmetres. Això és el que es coneix com a *Principi de Parsimònia*.

2.5 Anàlisi de residus

Els mètodes d'anàlisi de residus ens permeten veure si el model que hem ajustat és apropiat per les dades que tenim. En aquest apartat veurem les quatre gràfiques que dibuixen els paquets de R que utilitzem i que ens permeten analitzar els residus del model i així poder decidir si aquest és vàlid o no es pot acceptar.

Primer de tot tenim la gràfica dels residus en funció dels valors predits pel model. Aquesta gràfica ens permet veure si el model utilitza el tipus de relació adequada, com per exemple si el model hauria de ser no lineal enlloc de lineal. Aquest fet el podem observar si trobem tendències no aleatòries que segueixen els residus. Amb aquesta gràfica també podem comprovar si es compleix la homocedasticitat o, pel contrari, tenim problemes de dispersió irregular. Això es pot comprovar quan tenim un patró de dispersió no aleatori dels residus. Finalment, aquesta gràfica també ens permet trobar dades influents (outliers) que pertorbin el model.

La següent gràfica utilitza els *residus estandarditzats*, que es defineixen de la següent manera:

$$\text{Residus estandarditzats} = \frac{y_i - \hat{y}_i}{S\sqrt{1-h_{ii}}}, \textit{ on } h_{ii} = [X(X^tX)X^t]_{ii}$$

La segona gràfica que comentarem i la qual cal observar sempre que utilitzem models lineals és la *gràfica Q-Q*. Tenim que els residus han de seguir una distribució Normal. Aquesta gràfica ens permet comparar la distribució dels residus estandarditzats amb la distribució Normal teòrica, mitjançant els quantils. Per tant, cal comprovar que els residus estandarditzats segueixen aproximadament la línia recta diagonal que marca la distribució Normal teòrica i, en cas de que difereixin molt, no es podrà acceptar el model degut a la falta de normalitat dels residus.

La tercera gràfica que observem mostra l'arrel quadrada dels residus estandarditzats en funció dels valors predits pel model. En aquest cas, al igual que en la primera gràfica, cal que no s'observin tendències per així poder validar la hipòtesi d'homocedasticitat, és a dir, es vol observar una línia horitzontal amb els punts aleatòriament equiespaiats.

La última gràfica que s'observa correspon als residus estandarditzats en funció del *Leverage*, que es defineix com $h_{ii} = [X(X^tX)X^t]_{ii}$. Aquesta gràfica ens permet detectar observacions influents que caldria eliminar del model, ja que influeixen molt en el resultat d'aquest.

3. Metodologia: Models Lineals Generalitzats

En aquest capítol s'estenen els models lineals a través dels models lineals generalitzats. La informació d'aquest capítol s'ha extret dels llibres [1] i [2], així com dels apunts de l'assignatura Models Lineals i Lineals Generalitzats del màster MESIO UPC-UB.

3.1 Introducció

Els models lineals només ens permeten estudiar variables aleatòries que segueixen una distribució normal. A més a més, només ens permeten que l'esperança $\mu = \mathbb{E}[Y]$ sigui lineal amb les covariables. Per tant, per poder estudiar variables aleatòries que provenen d'altres distribucions, així com per abordar situacions en que l'esperança no sigui lineal amb les covariables, necessitem introduir els *models lineals generalitzats*, dels quals els models lineals explicats prèviament en són un cas particular.

A partir de l'equació (1) dels models lineals, es poden definir els models lineals generalitzats de la següent manera. Un model lineal generalitzat és aquell que admet una expressió de la forma:

$$\eta = g(\mu) = g(\mathbb{E}[Y]) = X\beta, \tag{4}$$

on primer de tot tenim la component aleatòria, que ve donada pel vector $Y^t = (Y_1, Y_2, \cdots, Y_n)$ i s'assumeix que

$$Y_i \sim exp\left(\frac{\theta_i y_i - b(\theta_i)}{a(\phi)} + c(y, \phi)\right) \tag{5}$$

També es té que Y_i és independent de Y_j $\forall i \neq j$.

A $X\beta$ se'l coneix com el *component determinista*, on X és una matriu n per p corresponent a les condicions experimentals i β és un vector de p components.

Finalment, η de (4) és la *funció d'enllaç* i pot ser qualsevol funció monòtona invertible.

En un model lineal generalitzat, el paràmetre θ_i que apareix en (5) s'anomena *paràmetre canònic* i el paràmetre ϕ s'anomena *paràmetre de dispersió*. Observis que el paràmetre canònic canvia al canviar les covariables (a l'igual que ho feia la μ dels models lineals). El paràmetre de dispersió, en canvi, és constant en totes les observacions, tal com ho era la σ^2 dels models lineals.

Definició 3.1. Es defineix la *funció d'enllaç canònica* com la funció η que satisfà $\eta = g(\mu) = \theta$.

Les funcions de densitat de les distribucions Normal, Binomial i Poisson es poden escriure com a la fórmula (5), i explicarem les propietats de cada una d'elles més endavant. Trobar la funció d'enllaç canònica de cada una d'elles serà important, ja que fa que $y^t X$ sigui un estadístic suficient, caldran menys iteracions per trobar l'estimador del vector β, com veurem més endavant en la Secció 3.3, i serà més fàcil d'interpretar el model.

Un cop presentat el model, ara anem a veure quines seran l'esperança i la variància de la variable resposta d'un model lineal generalitzat.

3.2 Funcions de versemblança, esperança i variància

Per calcular l'esperança i la variància de la variable aleatòria Y utilitzarem la funció de log-versemblança. Tenim per (5) que la funció de densitat de les components de Y són de la forma següent:

$$f(y_i;\theta_i,\phi) = exp\left(\frac{\theta_i y_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right)$$

Calculem ara la funció de log-versemblança, que és la següent:

$$l(\theta_i,\phi;y_i) = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi) \tag{6}$$

Tenim que si la funció de log-versemblança és integrable, té un únic màxim dins de l'espai de paràmetres, és de classe $\mathscr{C}^2$ i les seves derivades permeten l'intercanvi d'integrals amb derivades, aleshores es compleix que

$$\mathbb{E}\left[\frac{\partial l}{\partial\theta_i}\right] = 0, \;\; i = 1,...,p-1 \tag{7}$$

i també que

$$\mathbb{E}\left[\frac{\partial^2 l}{\partial\theta_i^2}\right] + \mathbb{E}\left[\frac{\partial l}{\partial\theta_i}\right]^2 = 0, \;\; i = 1,...,p-1 \tag{8}$$

Ara, aplicant (7) a la derivada de la funció de log-versemblança, que és $\frac{\partial l}{\partial\theta_i} = \frac{Y_i - b'(\theta_i)}{a(\phi)}$, obtenim que l'esperança val

$$\mathbb{E}[Y_i] = \mu_i = b'(\theta_i)$$

Calculem ara la variància utilitzant (8) i sabent que $\frac{\partial^2 l}{\partial\theta_i^2} = \frac{-b''(\theta_i)}{a(\phi)}$. Obtenim que $-\frac{b''(\theta_i)}{a(\phi)} + \frac{var(Y_i)}{a^2(\phi)} = 0$ i, per tant, la variància val

$$Var(Y_i) = b''(\theta_i)a(\phi)$$

Definició 3.2. A la part de la variància que depèn del paràmetre canònic θ_i, $b''(\theta_i)$ se l'anomena *funció de variància* i s'escriurà com $V(\mu_i)$.

Si tenim ara un vector de variables aleatòries $Y^t = (Y_1, Y_2, ..., Y_n)$ i sigui $y^t = (y_1, y_2, ..., y_n)$ una realització de Y. Aleshores, tenim que la funció de versemblança de Y és el producte de les funcions de densitat dels diferents elements de la realització y, és a dir,

$$L(\theta,\phi;y) = \prod_{i=1}^{n} f(y_i;\theta_i,\phi) = exp\left(\sum_{i=1}^{n}\frac{\theta_i y_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right)$$

on $\theta^t = (\theta_1, \theta_2, ..., \theta_n)$.

Calculem aleshores la funció de log-versemblança de Y, és a dir, apliquem el logaritme a la fórmula anterior i obtenim

$$l(\theta,\phi;y) = \sum_{i=1}^{n}\frac{\theta_i y_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi) \tag{9}$$

Aquesta funció la utilitzarem més endavant a la secció següent.

3.3 Càlcul de l'estimació dels paràmetres β

Volem trobar ara els estimadors màxim versemblants dels paràmetres β pel model definit a (4), un cop tenim que $y^t = (y_1, y_2, ..., y_n)$ és una realització del vector Y. Havíem trobat en (9) la funció log-versemblant, per tant, per calcular els estimadors, imposem ara que

$$0 = \frac{\partial l}{\partial \beta_j} = U_j = \sum_{i=1}^{N} \frac{(y_i - \mu_i)x_{ij}}{var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right), \ j = 1, \cdots, p-1$$

Utilitzant el mètode de Newton-Raphson per resoldre-ho, queda el següent mètode iteratiu per trobar l'estimador de β:

$$X^t W X b^{m+1} = X^t W Z, \tag{10}$$

on W és una matriu diagonal de mida n per n amb $w_{ii} = \frac{1}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$ i Z és un vector de n elements definit per

$$z_i = \sum_{j=1}^{p} x_{ij} b_j^m + (y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i}$$

L'esquema iteratiu que caldrà seguir per trobar l'estimador és el següent:

$$\beta^m \longrightarrow \eta \longrightarrow \mu \longrightarrow \theta \longrightarrow Var(Y) \longrightarrow \left\{ \begin{array}{l} W \\ Z \end{array} \right. \longrightarrow \beta^{m+1}$$

Per començar, prendrem de valor inicial μ_0 igual a la mitjana dels valors observats, començant el procés des del segon pas.

Observem en l'esquema que per calcular μ a partir de η sha de calcular la inversa de la funció g, cosa que pot resultar costosa. En canvi, si utilitzem la funció d'enllaç canònica definida a la Definició 3.1, aleshores passaríem directament de β^m a θ, estalviant-nos així el càlcul de la funció inversa. Per aquest motiu, és recomanable utilitzar la funció d'enllaç canònica d'entrada, i només quan aquesta no dóna bons resultats es considerarà una altra funció enllaç. Cal dir que (10) equival a fer uns mínims quadrats iteratius amb pesos. Els pesos són els elements de la diagonal de W.

Una vegada tenim calculats els valors dels paràmetres β, caldrà veure si el resultat és o no un bon model i poder comparar-lo amb d'altres.

En el cas d'un model lineal generalitzat, els valors predits es calcularan com:

$$\hat{\mu} = g^{-1}(X\hat{\beta})$$

3.4 Bondat d'ajust

Quan ja tenim el model amb els paràmetres estimats, voldrem ser capaços de decidir si és un bon model o, pel contrari, l'hem de rebutjar. Per fer això, es considerarà el *model complet* com aquell model que té tants paràmetres com observacions i que, per tant, donarà lloc a un ajust perfecte. Després es compararà el nostre ajust amb el del model complet mitjançant la següent definició:

Definició 3.3. Sigui $l(\hat{\mu}, \phi; y)$ el valor de la funció de log-versemblança del nostre model i sigui $l(y, \phi; y)$ el valor de la funció de log-versemblança del model complet. Es defineix la *deviança escalada* com

$$D^*(y; \hat{\mu}) = 2(l(y; y) - l(\hat{\mu}; y))$$

Per tal de comparar el nostre model amb el model complet, és a dir, per tal de fer el test d'hipòtesi:

$$\begin{cases} H_0 : \text{el nostre model} \\ H_1 : \text{model complet} \end{cases} \tag{11}$$

s'utilitzarà que sota la hipòtesi H_0, $D^*(y; \hat{\mu}) \sim \mathcal{X}^2_{n-p}$, on p és el nombre de paràmetres del nostre model. Per tant, en el nostre test a nivell α rebutjarem H_0 quan $D^*(y; \hat{\mu}) \geq \mathcal{X}^2_{1-\alpha, n-p}$. Observis que té sentit rebutjar H_0 quan la diferència de les dues log-versemblances sigui gran, o sigui quan $D^*(y; \hat{\mu})$ sigui gran.

A continuació veiem una figura[3] que mostra com funciona el test a nivell α per la distribució $\mathcal{X}^2$.

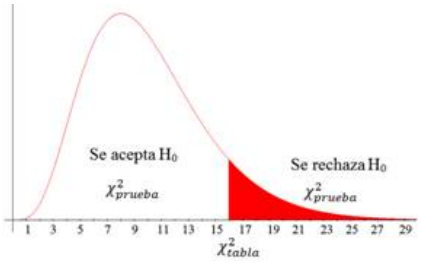


Figura 3: Gràfica de la zona de rebuig de la distribució Khi quadrat

A més a més, amb la deviança escalada es pot comparar dos models aniuats, tal com expliquem a continuació.

Definició 3.4. Dos models són aniuats si un conté tots els paràmetres que té l'altre i en té algun més.

Així doncs, si tenim dos models aniuats amb p_1 i p_2 paràmetres respectivament, $p_2 > p_1$ aleshores podem fer el següent test d'hipòtesis:

$$\begin{cases} H_0 : \text{model 1} \\ H_1 : \text{model 2} \end{cases}$$

comparant les deviances escalades. Això és degut a que sota la hipòtesi H_0, $D_1^* - D_2^* \sim \mathcal{X}^2_{p_2-p_1}$ i per tant rebutjarem H_0 del nostre test de nivell α quan $D_1^* - D_2^* \geq \mathcal{X}^2_{1-\alpha, p_2-p_1}$, ja que voldrà dir que el segon model aporta una significativa millor descripció de les dades.

Un altre mètode per testar si el nostre model ajusta correctament les dades és el mètode de Pearson, el qual consisteix a realitzar el test següent:

$$\begin{cases} H_0 : \text{Acceptem el model} \\ H_1 : \neg H_0 \end{cases} \tag{12}$$

[3] Gràfica extreta de *https://www.monografias.com/trabajos97/prueba-ji-cuadrado-excel-winstats-y-geogebra/prueba-ji-cuadrado-excel-winstats-y-geogebra.shtml*

Es defineix l'*estadístic generalitzat de Pearson* com:

$$X^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu_i})^2}{V(\hat{\mu_i})},$$

on $\hat{\mu_i}$ són els valors predits. El X^2 de Pearson és la suma de les discrepàncies existents entre valors observats i predits un cop estandarditzades. Per a portar a terme el test 12 amb l'estadístic de Pearson generalitzat hauríem de rebutjar el model quan $X^2 \geq \mathcal{X}^2_{1-\alpha,n-p}$.

Observem que l'estadístic de Pearson no ens permet comparar models aniuats, però és una mesura molt més intuïtiva.

Les tres seccions següents estan dedicades a veure quina forma prenen els models lineals generalitzats pels casos particulars de resposta Normal, Binomial i Poisson.

3.5 Distribució Normal

Per tal de veure que la distribució Normal es pot utilitzar com a distribució de la variable resposta en un model lineal generalitzat, hem de veure que podem expressar la funció de densitat de la distribució Normal com una *família exponencial* de la forma de (5). Manipulant una mica la densitat de la Normal tenim que aquesta es pot expressar de la següent forma:

$$f(y;\mu,\sigma^2) = exp\left(\frac{y\mu}{\sigma^2} - \frac{y^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} - log(\sqrt{2\pi\sigma^2})\right), \ \ \mu \in \mathbb{R}, \ \sigma^2 > 0 \tag{13}$$

Comparant (13) amb (5) podem trobar el valor de cadascun dels paràmetres esmentats anteriorment per la distribució Normal. A continuació, podem veure en la taula les principals característiques de la distribució Normal per als models lineals generalitzats:

	Distribució Normal
Paràmetre canònic	μ
Paràmetre de dispersió	σ^2
Enllaç canònic	$g(\mu) = \mu$
Funció variància	1
Deviança	$\sum_{i=1}^{n}(y_i - \hat{\mu_i})^2$
Residu de Pearson	$\sum_{i=1}^{n}(y_i - \hat{\mu_i})^2$

Taula 1: Taula de les característiques de la distribució Normal per MLG

Observis doncs, que els models lineals són el cas particular dels models lineals generalitzats en el que la distribució de la variable resposta és Normal i la funció enllaç és la identitat.

3.6 Distribució Binomial

Volem fer ara el mateix amb la distribució Binomial, és a dir, veure si podem expressar la seva funció de densitat de forma exponencial com a la fórmula (5). Manipulant les probabilitats de la Binomial, veiem

que queden de la forma següent:

$$f(y;p) = exp\left(ylog\left(\frac{p}{1-p}\right) + mlog(1-p) + log\binom{m}{y}\right), \quad p \in (0,1), \ y \in \{1,2,...,m\}, \tag{14}$$

on m és el nombre de rèpliques de la Binomial.

Igual que hem fet amb la distribució Normal, comparant (5) amb (14) veurem les característiques de la Binomial quan es considera distribució de la variable resposta en un model lineal generalitzat.

	Distribució Binomial
Paràmetre canònic	$\log\left(\frac{p}{1-p}\right)$
Paràmetre de dispersió	1
Enllaç canònic	$\log\left(\frac{\mu}{m-\mu}\right) = log\left(\frac{p}{1-p}\right)$
Funció variància	$\mu(1-\mu)$
Deviança	$2\sum_{i=1}^{n}\{y_i log(y_i/\hat{\mu}_i) + (m-y_i)log[(m-y_i)/(m-\hat{\mu}_i)]\}$
Residu de Pearson	$\sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{\hat{\mu}_i(1-\hat{\mu}_i)}$

Taula 2: Taula de les característiques de la distribució Binomial per MLG

3.7 Distribució Poisson

En aquesta secció farem el mateix que hem fet per la Normal i la Binomial en les seccions anteriors, però per la distribució de Poisson. La funció de probabilitats de la distribució de Poisson es pot expressar com en la fórmula (5) de la forma següent:

$$f(y;\lambda) = exp\left(ylog(\lambda) - \lambda - logy!\right), \ \ on\ \lambda > 0, \ y \in \mathbb{Z}^+ \cup \{0\} \tag{15}$$

Comparant (5) i (15) s'obté que les característiques de models lineals generalitzats corresponents a la distribució de Poisson són les de la taula següent:

	Distribució Poisson
Paràmetre canònic	$log(\lambda)$
Paràmetre de dispersió	1
Enllaç canònic	$log(\mu) = log(\lambda)$
Funció variància	μ
Deviança	$2\sum_{i=1}^{n}\{y_i log(y_i/\hat{\mu}_i) - (y_i - \hat{\mu}_i)\}$
Residu de Pearson	$\sum_{i=1}^{n}\frac{(y_i-\hat{\mu}_i)^2}{\hat{\mu}_i(1-\hat{\mu}_i)}$

Taula 3: Taula de les característiques de la distribució Poisson per MLG

Observis que per a la distribució Normal el paràmetre de dispersió coincideix amb la variància, que és constant i independent de les condicions experimentals. Pel cas de la Binomial i la Poisson, al ser uniparamètriques, el paràmetre de dispersió és constant igual a 1.

4. Descripció de la base de dades

Abans de veure les anàlisis realitzades als nois i noies del grau de matemàtiques, cal explicar primer quines són les dades que tenim i amb les quals treballarem.

La base de dades consta d'un fitxer amb 725 files corresponents als alumnes que han accedit al grau de matemàtiques entre els anys 2009 i 2019, ambdós inclosos, i amb 36 columnes que contenen les variables següents:

- **any_entrada:** És una variable categòrica que correspon a l'any d'entrada de l'alumne. Pren valors des de l'any 2009 al 2019.
- **CFIS:** És una variable categòrica que indica si l'alumne és CFIS o no. En la base de dades tenim que hi ha 270 alumnes CFIS i 455 que no ho són. Les seves categories són 0 pels no CFIS i 1 pels CFIS.
- **nota_acces:** És una variable numèrica contínua que correspon a la nota d'accés de l'alumne a la facultat. L'any 2009 la nota és sobre 10, la resta són sobre 14. Per aquest motiu, s'ha reescalat la nota dels alumnes que van entrar l'any 2009 per posar-la sobre 14.
- **nota_acces_min_curs:** És una variable numèrica contínua que indica la nota d'accés més baixa de l'any que va entrar l'alumne al grau. L'any 2009 la nota era sobre 10, la resta sobre 14. Per homogeneïtzar s'ha reescalat també la nota del 2009 per posar-la sobre 14.
- **provaNivell:** És una variable numèrica contínua que correspon a la nota que va treure l'alumne a la prova de nivell. No existeix nota de la prova de nivell per a tots els alumnes ja que és una prova opcional. De fet dels 725 alumnes hi ha 506 dels quals es disposa de la nota de la prova de nivell.
- **sexe:** És una variable categòrica que indica el gènere de l'alumne. A les dades hi ha 517 homes i 208 dones.
- **mitjana_FI:** És una variable numèrica contínua que correspon a la nota mitjana de la fase inicial de l'alumne. Aquesta variable només es té pels alumnes que han superat la fase inicial.
- **nombre_quad_FI:** És una variable discreta que indica el nombre de quadrimestres que va necessitar l'alumne per superar la fase inicial. A l'igual que la variable anterior, aquesta variable només existeix pels alumnes que han superat la fase inicial.
- **nota_Fonaments, nota_AlgLineal, ..., nota_ModelsMatematicsTecnologia:** Són variables numèriques contínues i es corresponen amb les notes de les 25 assignatures *obligatòries* del grau.
- **nombre_assigs_aprovades_1:** Indica el nombre d'assignatures que ha aprovat l'alumne a la primera. És una variable discreta que pren valors de 0 a 25.
- **graduat:** És una variable categòrica que indica si l'alumne s'ha graduat o no. Zero indica que no s'ha graduat i 1 que sí que s'ha graduat.

A més a més, hem hagut d'afegir algunes columnes a la base de dades per tal de poder fer les anàlisis, que són les següents:

- **nota_acces_14:** Com que l'any 2009 la nota de la selectivitat dels alumnes ponderava sobre 10 i la resta dels anys sobre 14, ha sigut necessari afegir una columna amb les notes d'accés escalades sobre 14.
- **nota_acces_min_curs_14:** Amb la nota d'accés mínima del curs tenim el mateix problema esmentat en l'anterior punt, per tant ha sigut necessari afegir una nova columna amb les notes de tall escalades sobre 14.
- **diferencia_notes:** És la diferència entre la nota d'accés de l'alumne i la nota d'accés mínima d'aquell any al grau. Ha sigut útil considerar-la en alguns dels models que hem realitzat. Aquesta variable mesura quin lloc ocupa un estudiant respecte als que han entrat en el seu mateix any.
- **grup:** És una variable categòrica que consta de 4 grups: els homes CFIS, les dones CFIS, els homes no CFIS i les dones no CFIS.
- **no_abandonar:** És una variable categòrica que indica si l'alumne ha abandonat el grau o no. Zero indica que ha abandonat i 1 que no ho ha fet. Per crear aquesta variable ha sigut necessari eliminar dels models on s'utilitzava els estudiants que van entrar els anys 2018 i 2019, ja que no teníem suficient informació per decidir si aquests alumnes havien abandonat o no. Dels estudiants que van entrar des de l'any 2009 fins al 2014 inclòs, hem considerat que l'alumne ha deixat el Grau de Matemàtiques si encara no s'ha graduat. Pels alumnes que van entrar l'any 2015 hem considerat que l'alumne ha abandonat si no s'ha matriculat de l'assignatura de Models Matemàtics de la Tecnologia, que és l'ultima assignatura obligatòria del grau i es realitza el primer quadrimestre del quart any. Pels alumnes que van entrar l'any 2016 hem considerat que l'alumne ha abandonat si no s'ha matriculat de l'assignatura de Teoria de la Probabilitat, la qual es realitza al primer quadrimestre del tercer any. Finalment, pels alumnes que van entrar l'any 2017 es considera que han abandonat si encara no han superat la fase inicial. En algun cas ha sigut necessari observar individualment els resultats de les assignatures per decidir si l'alumne ha abandonat el grau o no. A més a més, eliminem a 9 alumnes de la base de dades dels quals dubtem de la seva continuïtat al grau, quedant un total de 585 observacions.
- Finalment, també ha sigut necessari afegir columnes amb el logaritme o el quadrat d'algunes de les variables per utilitzar-les en les anàlisis. Això ha estat necessari quan els models han requerit de transformacions.

Per acabar, en gairebé totes les anàlisis hem hagut de reduir el nombre d'alumnes que utilitzàvem en la mostra ja que no complien algun dels requisits que necessitàvem. Quan tenim en compte la nota de la prova de nivell, hem hagut de suprimir tots aquells alumnes que no l'havien realitzat, quedant 506 dels 725 alumnes que tenim en total. A més a més, quan utilitzem la nota mitjana de la fase inicial, hem hagut d'eliminar les dades de tots aquells alumnes que no l'han superat, quedant un total de 494 alumnes que han acabat la fase inicial.

5. Comparació dels estudiants per gènere al entrar a la universitat

En aquest apartat volem comprovar si hi ha diferència en la capacitat en l'àmbit de les matemàtiques entre els nois i les noies quan entren a la universitat. Ho farem comparant la *nota d'accés* (Secció 5.1) i la nota de la *prova de nivell* (Secció 5.2) dels nois i de les noies. Observis que aquestes són dues variables en les quals els coneixements adquirits a la universitat no influeixen i venen donades únicament per l'aprenentatge previ a les escoles i instituts per part dels alumnes.

5.1 Comparació de la nota d'accés

Primer de tot volem veure si hi ha diferència entre nois i noies en quant a la nota amb la que accedeixen a la universitat, per veure si comencen els estudis de matemàtiques amb uns coneixements generals semblants. Per fer-ho, generem un diagrama de caixa de la nota d'accés en funció del gènere i així tenir una idea de si les notes entre homes i dones són semblants i si tenim observacions atípiques que poden influir en els resultats.

A la Figura 4 observem que sembla que les notes d'accés de nois i noies a la facultat són molt semblants, tant pel que fa als valors centrals com el que fa a la dispersió. També veiem que hi ha una dada en dones i dos en homes molt diferents de la resta, les quals procedim a eliminar de les dades per seguir amb l'anàlisi descriptiu, ja que molt probablement aquests estudiants hagin entrat per una via diferent de la de les proves d'accés a la universitat (PAU).

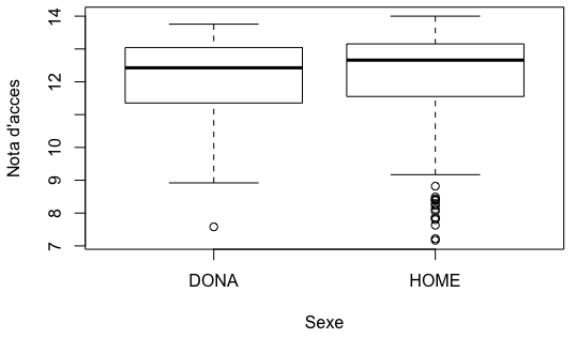


Figura 4: Diagrama de caixes de la variable *nota d'accés* per a cadascun dels gèneres

Per fer-ho, calculem la mitjana i la desviació estàndard de les notes d'accés per homes i per dones, les quals podem veure en la Taula 4.

	Homes	**Dones**
Mitjana	12.227858	12.155048
Desviació	1.257709	1.104162

Taula 4: Mitjana i desviació estàndard de la variable *nota d'accés* segons el gènere

Observem que tant la mitjana com la desviació estàndard són semblants per homes i per dones.

Una vegada tenim les dades amb les que treballarem, procedim a comprovar si les dades segueixen una distribució normal. A continuació fem un histograma de les notes comparant-les amb una distribució normal que pren per μ i σ^2 els valors de la mitjana aritmètica i la variància mostral per a cada gènere, els quals trobem tot seguit.

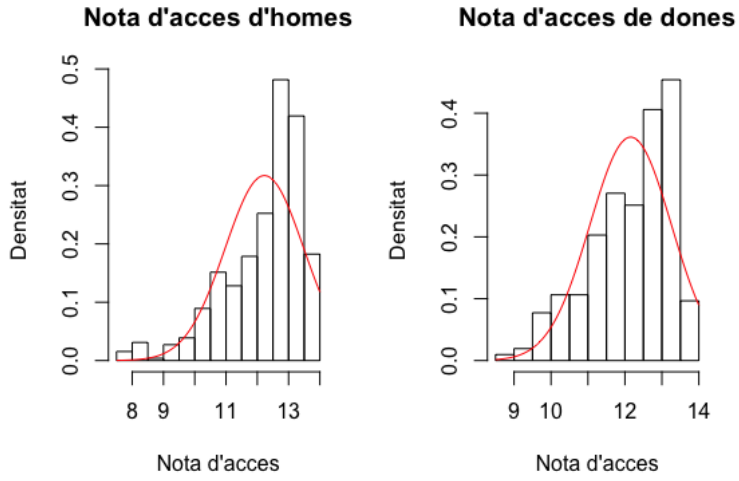


Figura 5: Histogrames de la variable *nota d'accés* segons el gènere

Observem que la normalitat és dubtosa degut a la falta de simetria de les dades, ja que hi ha molts alumnes amb notes molt altes, vora els 13 punts. Aquest resultat era d'esperar, ja que els estudiants que entren a la facultat no representen la població d'estudiants universitaris sinó només aquells que tenen una molt bona puntuació, d'aquí que hi hagi freqüències tant altes a puntuacions grans. Tot i així, assumirem normalitat en la *nota d'accés* per tal de poder aplicar els tests clàssics de comparació de dos valors esperats i de dues variàncies.

Ara volem saber si els valors esperats de la *nota d'accés* són iguals pels nois tant com per les noies. És a dir, assumirem que les variables *nota d'accés pels nois* (Y_1) i *nota d'accés per les noies* (Y_2) són Normals i volem comparar:

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases} \tag{16}$$

on μ_1 i μ_2 són, respectivament, iguals a $\mathbb{E}[Y_1]$ i $\mathbb{E}[Y_2]$.

Per veure-ho, comparem primer les variàncies de nois i de noies fent un F-test. Això ens permetrà saber

si el test de comparació de mitjanes l'hem de fer amb variàncies iguals o diferents. Per tant, denotant per σ_1^2 la variància de Y_1 i per σ_2^2 la variància de Y_2, el que volem comparar és:

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases} \tag{17}$$

El valor de l'estadístic de prova del test és $F = 1.2975$, amb 514 i 206 graus de llibertat. El resultat és que el p-valor del test és menor al nivell de significació 0,05 i, per tant, rebutgem la hipòtesi nul·la de que les variàncies són iguals, és a dir, les variàncies de nois i noies no són significativament iguals.

Per tant, ara per comparar els valors esperats entre nois i noies fem el t-test (16) assumint variàncies diferents. El valor de l'estadístic de prova és $t = -0.76914$ amb 430 graus de llibertat i el resultat és que el p-valor del test és major al nivell de significació 0.05. Això vol dir que el valor esperat de la *nota d'accés* de nois i noies no és significativament diferent. Podem concloure per tant que no s'observen diferències en la *nota d'accés* entre els nois i les noies.

5.2 Comparació de la nota de la prova de nivell

A continuació volem fer una anàlisi per saber si els nois i les noies quan entren a la facultat treuen una nota semblant a la *prova de nivell* que es realitza a la universitat abans de començar les classes. Així podrem veure si els nois i noies tenen un mateix nivell de matemàtiques en el moment que entren a la facultat. La comparació de la nota treta en la *prova de nivell* entre els nois i les noies té l'avantatge de ser una prova única per a tots i aquesta sí que s'espera que discrimini entre estudiants.

Per fer-ho, cal eliminar de les dades tots aquells alumnes que no han realitzat la *prova de nivell*, ja que aquesta no és obligatòria, quedant un total de 506 alumnes dels 725 dels quals tenim dades. A continuació generem un diagrama de caixa de la nota de la *prova de nivell* en funció del gènere per tenir una idea dels resultats i per veure si tenim observacions atípiques que calgui eliminar (veure Figura 6).

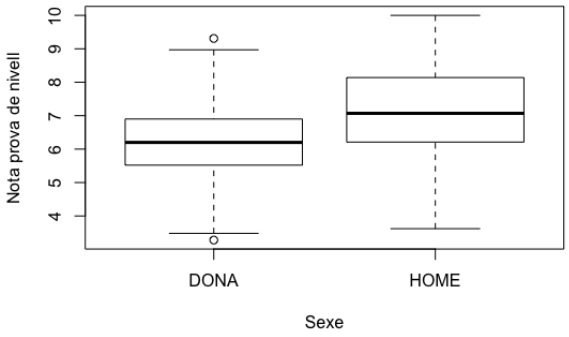


Figura 6: Diagrama de caixes de la nota de la *prova de nivell* per a cadascun dels gèneres

Observem que no hi ha observacions atípiques que es puguin eliminar i que sembla que els nois treuen millor nota que les noies en la *prova de nivell*, ja que la mediana dels nois és 7.07 i la de les noies és igual a 6.2. També observem més variabilitat en els nois ja que el rang interquartílic ($Q_3 - Q_1$) és superior en els nois que en les noies.

Una vegada que sabem amb quines dades hem de treballar, procedim a comprovar si les notes de la *prova de nivell* de nois i noies segueixen una distribució normal. Per fer-ho, calculem la mitjana i la desviació estàndard dels dos grups i fem un histograma per cada gènere. Veiem primer en la Taula 5 la mitjana i la desviació estàndard de les notes de la *prova de nivell* de nois i noies.

	Homes	**Dones**
Mitjana	7.105339	6.157126
Desviació	1.312225	1.174712

Taula 5: Mitjana i desviació estàndard de la nota de la *prova de nivell*

Observem que la mitjana de la nota de la *prova de nivell* és quasi 1 punt superior en els nois que en les noies i la desviació estàndard també és lleugerament superior pels nois, però molt semblant a la de les noies. A continuació, comparem l'histograma de les notes de la *prova de nivell* per a cada gènere amb una distribució normal que pren per μ i σ^2 la mitjana aritmètica i la variància mostral de cada un dels dos gèneres (veure Figura 7).

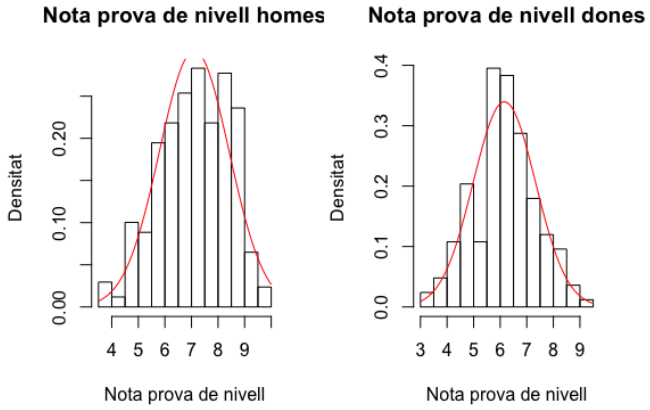


Figura 7: Histogrames de la nota de la *prova de nivell* segons el gènere

Observem en els histogrames que sembla que la nota de la *prova de nivell* segueix una distribució normal, per tant clarament la *prova de nivell* "normalitza"els estudiants en el sentit de que n'hi ha amb notes baixes, cosa que no passava amb la *nota d'accés*. Assumint ara que les notes de la *prova de nivell* segueixen una distribució normal, volem comprovar si els nois i les noies treuen una nota semblant. Per fer-ho, mirem primer si la variància pels dos gèneres és la mateixa amb un F-test, és a dir, realitzem el test (17) per a la nota de la *prova de nivell*. El valor de l'estadístic de prova és $F = 1.2478$, amb 338 i 166 graus de llibertat. El resultat és que el p-valor és major al nivell de significació 0.05 i, per tant, podem suposar que les variàncies no són significativament diferents.

Llavors fem un t-test per comparar el valor esperat de la nota de la *prova de nivell* de nois i noies assumint variàncies iguals. És a dir, realitzem el test (16) però ara assumint variàncies iguals. El valor de l'estadístic de prova és $t = -8.2089$ amb 365 graus de llibertat i el p-valor del t-test és inferior al nivell de significació 0.05, és a dir, tenim que la nota de la *prova de nivell* no és significativament igual per nois i per noies. Concloem finalment que els nois treuen una nota superior a les noies en la *prova de nivell*.

6. Anàlisi de la nota mitjana de la fase inicial en funció de la nota d'accés i la nota d'accés mínima del curs

Una vegada hem analitzat com entren els estudiants al grau de matemàtiques, volem analitzar els seus resultats una vegada superada la fase inicial. Per tant, en aquesta secció procurarem veure si hi ha una diferència significativa en la variable nota *mitjana de la fase inicial* entre els nois i les noies. Tanmateix, també estem interessats en saber quines són les altres variables que influeixen en la nota i de quina forma ho fan. La manera de fer-ho és creant un model que expliqui de la millor manera possible la *nota mitjana de la fase inicial* dels alumnes, amb la informació que tenim quan entren a la facultat. En aquest apartat treballarem amb les variables explicatives *gènere, CFIS, nota d'accés, nota d'accés mínima del curs, diferència de les dues notes* i *any d'entrada*.

El primer pas és eliminar de la base de dades tots aquells estudiants que no han finalitzat la fase inicial, quedant un total de 494 alumnes. Hem decidit no treballar amb la nota de la *prova de nivell* degut a que en aquest apartat procurarem trobar un model amb la informació que tenim dels alumnes que no depèn de la facultat. A més a més, hauríem d'eliminar més alumnes de la base de dades, així que deixarem pel següent apartat un model que tingui en compte la nota de la *prova de nivell*.

Comencem fent estadística descriptiva i en particular fent un diagrama de caixa de la nota *mitjana de la fase inicial* en funció del gènere per tenir una idea de les diferències i veure si tenim observacions atípiques.

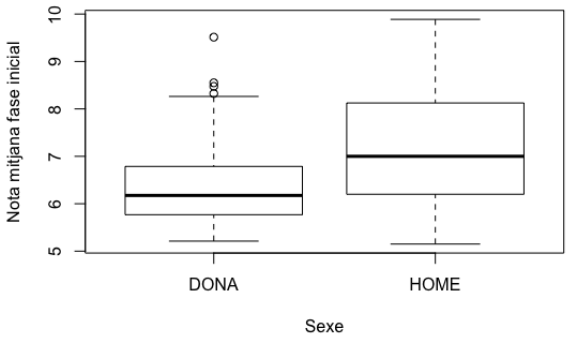


Figura 8: Diagrama de caixes de la *nota mitjana de la fase inicial* per a cadascun dels gèneres

De la Figura 15 observem que sembla ser que els nois treuen una nota superior a les noies, exactament la diferència entre la mediana dels homes i la de les dones és de 0.825 punts, però això no vol dir que el gènere sigui influent en el model que obtinguem de la nota *mitjana de la fase inicial*. No hi ha observacions que puguem eliminar pels homes i per les dones si que n'hi ha una que sembla ser que ha tret una nota molt superior a la resta, però hem decidit deixar-la. De la Figura 15 també observem que la distància interquartílica associada als homes és superior a la de les dones, la qual cosa indica que les notes dels homes varien més que les de les dones.

Ara que ja tenim les dades amb les que treballarem, anem a dibuixar diferents gràfiques que ens ajudin

a tenir una idea de com afecten les diferents variables amb les que treballarem a la nota *mitjana de la fase inicial*. Les primeres dues gràfiques que realitzem consisteixen en la nota *mitjana de la fase inicial* depenent del gènere i de la *diferència de notes* en el primer cas i de la *nota d'accés* en el segon cas.

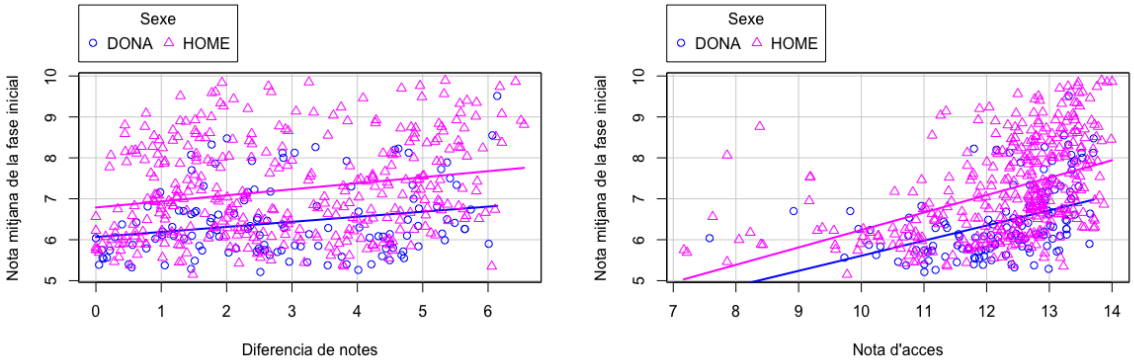


Figura 9: Gràfiques de la nota *mitjana de la fase inicial* en funció del gènere i de la *diferència de notes* (esquerra) i de la *nota d'accés* (dreta)

En les dues gràfiques de la Figura 9 es veu que les notes dels nois estan lleugerament per sobre de les notes de les noies i que les rectes de regressió són paral·leles. Això vol dir que l'augment d'una unitat de la *nota d'accés* o bé de la *diferència de notes* afecta de la mateixa manera a la mitjana de la fase inicial a nois i a noies.

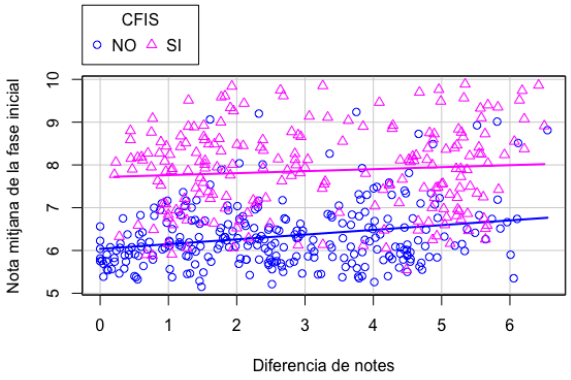


Figura 10: Gràfica de la nota *mitjana de la fase inicial* en funció de la *diferència de notes* i de si l'alumne és *CFIS* o no

També realitzem una gràfica de la nota *mitjana de la fase inicial* en funció de la *diferència de notes* i de si l'alumne és *CFIS* o no, i observem a la Figura 10 que la nota de la fase inicial dels estudiants *CFIS* està clarament per sobre de la dels alumnes que no ho són, la qual cosa ja era d'esperar. Les rectes de regressió corresponents també són pràcticament paral·leles, indicant que l'augment d'una unitat de la variable *diferència de notes* repercuteix de forma semblant en l'increment de la nota *mitjana de la fase inicial* tant en els *CFIS* com en els no *CFIS*.

Finalment, realitzem una gràfica de la nota *mitjana de la fase inicial* en funció de la *diferència de notes*

i ho farem per a cadascun dels gèneres pels *CFIS* i per cadascun dels gèneres pels no *CFIS*, mitjançant la variable *grup*.

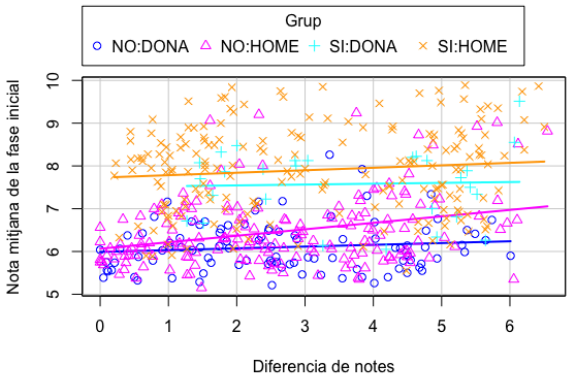


Figura 11: Gràfica de la nota *mitjana de la fase inicial* en funció de la *diferència de notes* i del *grup*

En aquesta gràfica observem que hi ha més diferència en la nota *mitjana de la fase inicial* depenent de si és *CFIS* o no que depenent del gènere.

Atès que trobar un model per a unes dades és una cosa dinàmica, explicarem per sobre els passos realitzats en els models intermedis i explicarem amb més detall només el model final. A continuació, realitzem uns models senzills, amb poques variables, per veure quines variables van sortint que influeixen més en la nota *mitjana de la fase inicial*. El primer model que ajustem consisteix en explicar la nota *mitjana de la fase inicial* en funció del gènere i de la *diferència de notes*. El resultat és que les dues variables són influents, però trobem que és un model molt pobre, ja que només explica un 13% de la variabilitat de la nota mitjana de la fase inicial. Fem el mateix model només pels estudiants *CFIS* i un altre només per estudiants no *CFIS*, però els models surten encara pitjor. Decidim aleshores fer un nou model afegint la variable CFIS i surt de nou que totes les variables són influents, sent aquesta última variable la que més error redueix en el model. El resultat és que aquest nou model explica un 46% de la variabilitat de la resposta, és a dir, de la variabilitat existent a la nota *mitjana de la fase inicial*, gairebé la meitat ve explicada pel gènere, per si ets *CFIS* o no i per la *diferència de notes*.

Intentem ara afegir interaccions de primer ordre entre les variables que tenim, per veure si això ens permet millorar el model i per veure si la manera en que una variable categòrica influeix a la resposta depèn de l'altre variable amb la qual es combina. Afegim primer la variable *sexe*CFIS*, però ens surt en el model que aquesta nova variable que hem afegit no és influent. Això vol dir que el gènere afecta de la mateixa manera a la nota *mitjana de la fase inicial* tan si la persona és *CFIS* com si no ho és. Provem ara a afegir com a variable la interacció *sexe*diferència de notes*, però, de nou, aquesta nova variable no és influent. És un resultat que esperàvem, ja que com havíem vist a la Figura 9 el pendent de les rectes per homes i per dones és el mateix. Finalment, provem d'afegir la variable *CFIS*diferència de notes*, però el resultat torna a ser que aquesta nova variable no és influent. Això era d'esperar pel que havíem vist a la Figura 10, ja que les rectes per *CFIS* i no *CFIS* eren també quasi paral·leles. Per tant, tenim que afegint interaccions de primer ordre el nostre model no millora, ja que aquestes són no significatives.

Provem ara el model que cont la *nota d'accés* i la *nota d'accés mínima del curs* en comptes d'utilitzar la diferència entre aquestes dues notes i el comparem amb el model anterior. En aquest nou model totes les variables són significatives, l'estadístic R^2 de bondat d'ajust ha passat d'un 46% a un 49% i reduïm molt

lleugerament l'error residual, ja que aquest en l'anterior model era igual a $\hat{\sigma} = 0.8467$ i ara és $\hat{\sigma} = 0.8245$. Com que les gràfiques de normalitat i de residus són molt semblants, ens quedem amb aquest segon model com el millor fins ara. Provem, igual que hem fet abans, d'afegir noves variables que siguin interaccions de primer ordre entre les variables que tenim però, novament, aquestes noves variables no són influents.

Per intentar trobar un model que expliqui més d'un 50% de la variabilitat en la resposta, intentem afegir noves variables als dos models bons que tenim fins ara (l'últim que hem esmentat i el que utilitza la *diferència de notes* en comptes de la *nota d'accés* i la *nota d'accés mínima dels curs*). Provem d'afegir en els dos models la variable *any d'entrada* com a factor i el millor model dels dos nous que obtenim és el que utilitza la variable *diferència de notes*. Per tant, comparem ara aquest darrer model amb el que teníem que era millor sense any d'entrada. Al afegir la variable *any d'entrada*, augmenta lleugerament el R^2 ajustat i es redueix també molt lleugerament l'error residual, mentre que les gràfiques de normalitat i de residus són molt semblants. Al donar uns resultats tan semblants, decidim quedar-nos amb el model que no té en compte *l'any d'entrada*, ja que incloent-lo estem afegint una altra variable i molts més paràmetres (un per cada any) i, per tant, complicant molt més el model. A més a més, si volem predir la nota *mitjana de la fase inicial* dels alumnes de nou ingrés a la facultat amb la informació que tenim d'ells quan entren a la facultat, necessitem no tenir en compte l'*any d'entrada*.

A vegades ajustar un model lineal havent prèviament transformat alguna de les variables dóna bons resultats. Malgrat que el model que tenim en aquest estat ja el considerem prou bo, mirarem si aplicant logaritmes a la variable resposta i també a la variable *nota d'accés* aquest millora, en el sentit de que l'R^2 augmenta i els residus queden millor. El resultat és que totes les variables són significatives i el R^2 ajustat és molt semblant al model que ja teníem, per tant, no ens surt a compte complicar més el model utilitzant els logaritmes de les variables, ja que la interpretació en escala logarítmica és molt més complicada.

Finalment, provem d'afegir la *nota d'accés al quadrat* com a variable. El resultat d'aquest nou model és que totes les variables surten significatives i tenim el millor R^2 ajustat amb un valor de 0.51, però tenim multicolinealitat entre les variables *nota d'accés* i *nota d'accés al quadrat*, ja que aquestes tenen un VIF de 150.75 i 153.6 respectivament. Per eliminar la colinealitat cal suprimir una de les dues variables. Suprimint la variable *nota d'accés*, el resultat és un model molt semblant al inicial abans d'afegir la nova variable i, per tant, ens quedem amb el primer degut a que és més senzill d'interpretar.

Per tant doncs, el model final és:

$$Y_i = \beta_0 + \beta_1 \text{Nota d'accés} + \beta_2 \text{Nota d'accés mínima} + \beta_3 \delta_{G\text{ènere}} + \beta_4 \delta_{CFIS} + e_i, \tag{18}$$

on $\delta_{g\text{ènere}}$ és igual a 1 pels homes i 0 per les dones i δ_{CFIS} és igual a 1 pels estudiants *CFIS* i 0 pels no *CFIS*.

A continuació, adjuntem la informació més rellevant del model final utilitzant la informació que es té dels alumnes quan entren a la facultat, mitjançant les funcions summary (Figura 12) i Anova (Figura 13) de R.

Podem observar a la Figura 12 que amb les variables que utilitzem en el model (18), les quals són totes influents, expliquem gairebé un 50% de la variabilitat de la nota *mitjana de la fase inicial* i tenim que l'error estàndard residual és de $\hat{\sigma} = 0.8245$. El fet que el nostre model no expliqui més del 50% de la variabilitat existent en la variable resposta té sentit, perquè hi ha variables com les hores d'estudi d'un estudiant que afecten considerablement la nota de la fase inicial, i a aquestes variables no hi tenim accés. Observis també que rebutgem el model nul respecte el model (18), ja que realitzant el test d'Omnibus (3) obtenim que el valor de l'estadístic de prova és $F = 119.5$ amb 4 i 489 graus de llibertat i el p-valor val $2.2 \cdot 10^{-16}$, inferior al nivell de significació 0.05 i rebutgem, per tant, la hipòtesi de que tots els coeficients

són nuls. També tenim que no hi ha multicolinealitat en el model, ja que el VIF de totes les variables és proper a 1. A més a més, a la Figura 13 podem observar que la variable que més influeix a l'hora de reduir l'error residual és si l'alumne és *CFIS* o no, amb una gran diferència respecte la *nota d'accés* i el gènere.

```
Call:
lm(formula = dades$mitjana_FI ~ dades$sexe + dades$nota_acces_14 +
    dades$nota_acces_min_curs_14 + dades$CFIS, data = dades)

Residuals:
     Min       1Q   Median       3Q      Max
-2.30136 -0.56595 -0.04354  0.48841  2.72665

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    3.73134    0.41857   8.914  < 2e-16 ***
dades$sexeHOME                 0.40600    0.08820   4.603 5.32e-06 ***
dades$nota_acces_14            0.23789    0.03642   6.531 1.64e-10 ***
dades$nota_acces_min_curs_14 -0.05140    0.02245  -2.290   0.0225 *
dades$CFISSI                   1.22445    0.08533  14.350  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8245 on 489 degrees of freedom
Multiple R-squared:  0.4944, Adjusted R-squared:  0.4903
F-statistic: 119.5 on 4 and 489 DF,  p-value: < 2.2e-16
```

Figura 12: Resum del model (18)

```
Anova Table (Type II tests)

Response: dades$mitjana_FI
                              Sum Sq  Df F value    Pr(>F)
dades$sexe                     14.40   1  21.187 5.316e-06 ***
dades$nota_acces_14            29.00   1  42.657 1.638e-10 ***
dades$nota_acces_min_curs_14    3.56   1   5.242   0.02247 *
dades$CFIS                    139.99   1 205.920 < 2.2e-16 ***
Residuals                     332.44 489
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 13: Funció Anova del model (18)

Per veure que podem considerar aquest model com a un bon model, adjuntem també la gràfica de residus contra valors predits i la gràfica Q-Q per analitzar la normalitat dels residus, i comprovar les hipòtesis d'independència i d'igualtat de variàncies.

De la Figura 14 part esquerra veiem que hi ha clarament dues poblacions de residus, aquells associats als estudiants *CFIS* que tenen un valor predit de la nota *mitjana de la fase inicial* clarament superior a aquells associats als estudiants que no ho són. No veiem patrons, i tampoc veiem que no es pugui acceptar la hipòtesi d'igualtat de variàncies. De la gràfica que apareix a la part dreta de la Figura 14 deduïm que els residus poden assumir-se normals ja que s'ajusten prou bé a la recta.

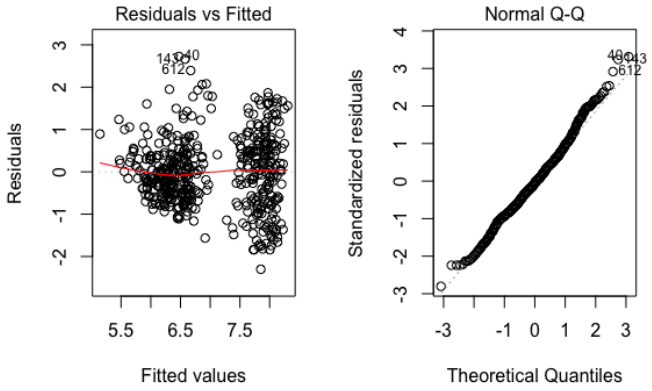


Figura 14: Gràfiques de residus del model (18)

A continuació volem veure quins són els valors predits pel model (18) de la nota *mitjana de la fase inicial* dels estudiants que van entrar al Grau de Matemàtiques l'any 2019, que no són *CFIS*, que tenen una *nota d'accés* mitja per aquell any de 13.13 i la *nota d'accés mínima del curs* 2019/2020 que és de 10.548. Els resultats per homes i dones són, respectivament

$$\text{Home no CFIS:} \quad \hat{y} = 4.136 + 0.24 \cdot 13.13 - 0.05 \cdot 10.548 = 6.72$$

$$\text{Dona no CFIS:} \quad \hat{y} = 3.731 + 0.24 \cdot 13.13 - 0.05 \cdot 10.548 = 6.31$$

Ara fem el mateix per estudiants *CFIS* i els resultats per homes i dones són, respectivament

$$\text{Home CFIS:} \quad \hat{y} = 5.36 + 0.24 \cdot 13.13 - 0.05 \cdot 10.548 = 7.94$$

$$\text{Dona CFIS:} \quad \hat{y} = 4.95 + 0.24 \cdot 13.13 - 0.05 \cdot 10.548 = 7.54$$

Finalment, del model (18) podem concloure que els nois que entren a la facultat treuen 0.4 punts més de nota *mitjana de la fase inicial* que les noies que han entrat el mateix any (tenint la mateixa *nota d'accés mínima del curs*), tenint la mateixa *nota d'accés* i estant dintre del mateix grup (són tots dos *CFIS* o no ho són). Tot i així, observis que no és comparable amb la diferència entre la nota *mitjana de fase inicial* esperada per un estudiant *CFIS* i un que no ho és, ja que amb les mateixes condicions els estudiants *CFIS* treuen 1.2 punts més de mitjana de la fase inicial que els que no ho són.

7. Anàlisi de la nota mitjana de la fase inicial en funció de la nota de la prova de nivell

En aquesta secció volem veure com influeix la nota de la *prova de nivell* a la nota *mitjana de la fase inicial*, així com trobar el millor model que expliqui la nota de la fase inicial i quines variables li afecten. Per fer-ho, treballarem amb totes les variables que hem utilitzat en la secció anterior més la nota de la *prova de nivell*.

Per començar, cal eliminar de la base de dades tots aquells alumnes que no hagin superat la fase inicial i també tots aquells que no hagin realitzat la prova de nivell, quedant un total de 338 persones per analitzar. Aleshores fem un diagrama de caixa de la *mitjana de la fase inicial* en funció del gènere per tenir una idea del comportament de la variable i per veure si tenim observacions atípiques.

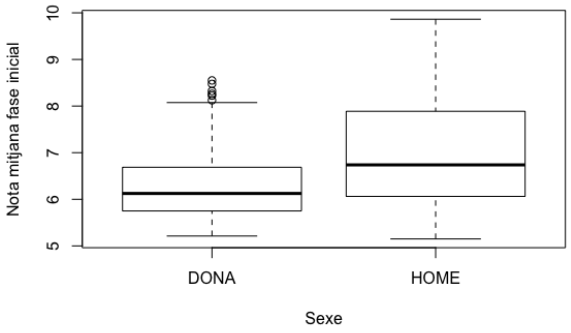


Figura 15: Diagrama de caixes de la nota *mitjana de la fase inicial* per a cadascun dels gèneres

Observem que també ara la *nota mitjana* dels nois de la fase inicial està una mica per sobre que la de les noies. Concretament la mediana dels nois és 6.738 i la de les noies 6.125. També ara la dispersió obtinguda amb els nois és superior a la de les noies, tal com cabia esperar perquè estem treballant amb una submostra de les dades considerades a la Secció 6. També veiem que no tenim observacions per eliminar.

Realitzem a continuació una gràfica de la nota *mitjana de la fase inicial* en funció de la *prova de nivell* i del gènere a veure quina influència té. Aquest és un punt totalment descriptiu.

Podem observar a la Figura 16 que les rectes de regressió per nois i noies es creuen, per tant, caldrà comprovar si la interacció *sexe*prova de nivell* és influent. Curiós i important observar que les noies que tenen una nota de la *prova de nivell* més baixa tenen una nota de la *mitjana de la fase inicial* més alta que els nois que tenen una mateixa nota de la *prova de nivell*. Això vol dir que les noies lluny de desanimar-se al treure una nota baixa a la *prova de nivell*, s'esforcen per a augmentar el seu rendiment. Finalment, es pot observar a la gràfica que el núvol de punts té una certa tendència de creixement que no sembla lineal, cosa que s'haurà de tenir en compte a l'hora de realitzar els models, ja que potser explicarà millor el *quadrat de la prova de nivell* que no la nota de la *prova de nivell*, cosa que comprovarem.

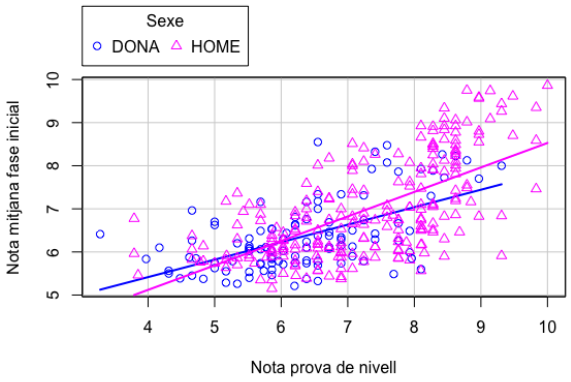


Figura 16: Gràfica de la nota *mitjana de la fase inicial* en funció de la nota de la *prova de nivell* i del gènere

Al igual que a la secció anterior, començarem amb models més senzills i anirem posant més variables per veure si aconseguim millorar la bondat d'ajust del model. Comencem amb un model que analitza la nota *mitjana de la fase inicial* en funció del gènere, de la nota de la *prova de nivell* i de *l'any d'entrada* (ja que pot influir en la nota de la prova de nivell al canviar la prova cada any). En aquest model surt que el gènere no és influent, aleshores eliminem la variable gènere i ens queda un model que té com a variables explicatives *l'any d'entrada* i la nota de la *prova de nivell* i que té un R^2 ajustat de 0.45. Veient les gràfiques dels residus observem que els residus poden assumir-se normals, però la gràfica de valors predits contra els residus mostra una lleugera tendència parabòlica, que podria estar causada per la falta d'algun terme d'ordre major en el model.

Anem a intentar explicar més variabilitat de la nota mitjana de la fase inicial afegint variables. En el següent model tornem a utilitzar les variables del primer model i hi afegim les variables *CFIS* i la diferència entre la *nota d'accés* i la *nota d'accés mínima del curs*. Aquest nou model té totes les variables influents (inclòs el gènere) i tenim que el R^2 ajustat és de 0.61, molt millor que el de l'anterior. Havíem vist anteriorment a la gràfica que semblava que hi havia interacció entre les variables gènere i nota de la *prova de nivell*, per tant, afegim al model la variable *sexe*prova de nivell*. El resultat és que aquesta nova variable no és influent, així que ens quedem amb el model sense la interacció.

Procedim ara a eliminar la variable *any d'entrada*, ja que ens complica el model afegint molts paràmetres i voldríem un model que no tingui en compte *l'any d'entrada*, per així ser capaços de predir la nota *mitjana de la fase inicial* dels nous alumnes, una vegada tenim la nota de la *prova de nivell*. El resultat obtingut al treure aquesta variable és que ens surt un model amb el R^2 ajustat lleugerament inferior, les gràfiques de residus molt semblants i amb la variable gènere no significativa. Procedim aleshores a eliminar la variable gènere i el resultat és un model amb totes les variables significatives i un R^2 ajustat de 0.58, el qual és lleugerament més baix que el model que utilitzava *l'any d'entrada*, però molt més simple i fàcil d'interpretar. Com que el R^2 ajustat no varia gaire, les gràfiques de residus tampoc i aquest nou model utilitza molts menys paràmetres, optem per utilitzar aquest darrer model ja que és molt més senzill i explica una variabilitat semblant de la nota *mitjana de la fase inicial*. A més a més, sense *l'any d'entrada* podrem predir la nota *mitjana de la fase inicial* d'un alumne una vegada es sàpiga la seva nota de la *prova de nivell*. Provem de fer aquest darrer model però utilitzant la *nota d'accés* i la *nota d'accés mínima del curs* en comptes de la diferència entre aquestes dues notes i, encara que augmenta molt lleugerament el R^2 ajustat, utilitza dos variables més, ja que el gènere en aquest cas surt que és una variable significativa, i empitjora les gràfiques dels residus. Per tant, optem per descartar aquest darrer model.

Com hem comentat a la Secció 6, a vegades ajustar un model lineal havent transformat alguna de les variables dóna bons resultats. Per tant, provem de fer el mateix model que teníem però aplicant logaritmes a la nota *mitjana de la fase inicial* i a la nota de la *prova de nivell*, per veure si ajustem millor les dades. El resultat és un model amb gràfiques de residus semblants i un pitjor R^2 ajustat, a part de complicar el model afegint logaritmes, així que decidim descartar aquest nou model.

Havíem comentat que en la Figura 16 l'evolució de la nota *mitjana de la fase inicial* en funció de la nota de la *prova de nivell* no semblava tenir un creixement lineal. Per tant, provem a afegir la variable nota de la *prova de nivell al quadrat* per veure si aconseguim millorar el model, en el sentit de que l'R^2 augmenti i els residus quedin millor, tornant a afegir la variable gènere. El resultat és que el gènere torna a no ser influent, així que l'eliminem, donant lloc a un model que té multicolinealitat entre les variables *prova de nivell* i *prova de nivell al quadrat*. Observem utilitzant la funció Anova del R que la variable nota de la *prova de nivell al quadrat* redueix més l'error residual que la nota de la *prova de nivell* i provem a eliminar doncs aquesta última. Comparem llavors aquest nou model amb el que teníem abans d'afegir la nota de la *prova de nivell al quadrat* (que és el mateix model canviant la nota de la *prova de nivell* per la nota de la *prova de nivell al quadrat*). El resultat és que aquest nou model té un R^2 ajustat lleugerament superior i els residus surten millor, sense que hi hagi cap patró a la gràfica de valors predits contra els residus. Per tant, optem per quedar-nos amb aquest darrer model, que és el següent:

$$Y_i = \beta_0 + \beta_1 \text{Prova de nivell}^2 + \beta_2 \text{Diferència de notes} + \beta_3 \delta_{CFIS} + e_i, \tag{19}$$

on δ_{CFIS} és igual a 1 pels estudiants *CFIS* i zero pels no *CFIS*.

A continuació, adjuntem la informació més rellevant d'aquest model final, mitjançant les funcions summary (Figura 17) i Anova (Figura 18) de R.

```
Call:
lm(formula = dades$mitjana_FI ~ dades$provaNivell2 + dades$diferencia_notes +
    dades$CFIS, data = dades)

Residuals:
     Min       1Q   Median       3Q      Max
-1.97243 -0.44552 -0.05067  0.40581  2.10689

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)          5.118868   0.125661  40.735  < 2e-16 ***
dades$provaNivell2   0.023321   0.002603   8.959  < 2e-16 ***
dades$diferencia_notes 0.064875   0.022401   2.896  0.00403 **
dades$CFISSI         1.044782   0.103112  10.133  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6937 on 334 degrees of freedom
Multiple R-squared:  0.5939, Adjusted R-squared:  0.5902
F-statistic: 162.8 on 3 and 334 DF,  p-value: < 2.2e-16
```

Figura 17: Resum del model (19)

Podem observar a la Figura 17 que amb les variables que utilitzem en el model (19), les quals són totes influents, expliquem un 59% de la variabilitat de la nota *mitjana de la fase inicial* i tenim que l'error estàndard residual és de $\hat{\sigma} = 0.69$. Observis també que rebutgem el model nul respecte el model (19), ja que realitzant el test d'Omnibus (3) obtenim que l'estadístic de prova és $F = 162.8$ amb 3 i 334 graus de llibertat i el p-valor val $2.2 \cdot 10^{-16}$, inferior al nivell de significació 0.05 i rebutgem, per tant, la hipòtesi de que tots els coeficients són nuls. També tenim que no hi ha multicolinealitat, ja que el VIF de totes les variables és proper a 1. A més a més, a la Figura 18 podem observar que la variable que influeix més a l'hora de reduir l'error residual és si l'alumne és *CFIS* o no ho és (igual que en la Secció 6), seguit de la nota de la *prova de nivell*.

```
Anova Table (Type II tests)

Response: dades$mitjana_FI
                       Sum Sq  Df F value    Pr(>F)
dades$provaNivell2     38.625   1  80.257 < 2.2e-16 ***
dades$diferencia_notes  4.036   1   8.387  0.004028 **
dades$CFIS             49.412   1 102.668 < 2.2e-16 ***
Residuals             160.746 334
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figura 18: Funció Anova del model (19)

Per veure que podem considerar aquest model com a un bon model, adjuntem també la gràfica de residus contra valors predits i la gràfica Q-Q per analitzar la normalitat dels residus, i comprovar les hipòtesis d'independència i d'igualtat de variàncies.

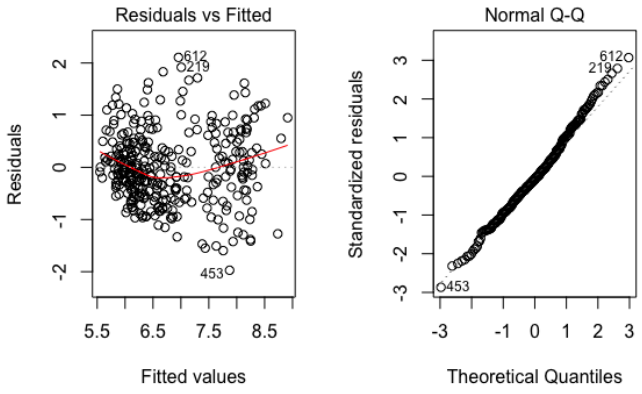


Figura 19: Gràfiques de residus del model 19

De la Figura 19 part esquerra observem que els residus no sembla que segueixin cap patró i tampoc veiem que no es pugui acceptar la hipòtesi d'igualtat de variàncies. De la gràfica Q-Q que apareix a la part dreta de la Figura 19 deduïm que els residus poden assumir-se normals ja que s'ajusten prou bé a la recta.

A continuació volem veure quins són els valors predits pel model (19) de la nota *mitjana de la fase inicial* dels estudiants que van entrar al Grau de Matemàtiques l'any 2019, que van realitzar la *prova de nivell* i van treure una nota mitja (7.25) i que tenen una *diferència de notes* mitja (2.63). Els resultats pels estudiants *CFIS* i els que no ho són els veiem a continuació:

$$\text{Estudiant CFIS:} \quad \hat{y} = 6.16 + 0.02 \cdot (7.25)^2 + 0.06 \cdot 2.63 = 7.56$$

$$\text{Estudiant no CFIS:} \quad \hat{y} = 5.12 + 0.02 \cdot (7.25)^2 + 0.06 \cdot 2.63 = 6.52$$

Finalment, d'aquest model podem concloure que amb la informació que es té dels alumnes després de realitzar la prova de nivell, el gènere no és una variable influent en la nota mitjana que tindran els alumnes en la fase inicial. Això pot ser degut a que, com havíem observat a la Secció 5.2, ja es tingui en compte la diferència segons el gènere en la nota de la *prova de nivell*. A més a més, la variable *CFIS* torna a tenir una

gran influència, ja que els alumnes que ho són treuen un punt més de mitjana de la fase inicial que els que no ho són. Finalment, veiem que la nota de la *prova de nivell* és bastant més influent que la diferència entre la nota d'accés i la nota de tall, és a dir, la prova específica de coneixements de matemàtiques realitzada a la facultat influeix més que les notes que porten els alumnes del batxillerat i de la selectivitat on s'avaluen també altres àmbits, cosa que era d'esperar.

8. Anàlisi del nombre de quadrimestres emprats en la fase inicial segons el gènere

L'objectiu d'aquesta secció és veure si hi ha diferència en el nombre de quadrimestres que necessiten els nois i les noies per superar la fase inicial. La fase inicial en els estudis del Grau de Matemàtiques està formada per les assignatures del primer curs que són un total de 8 (4 cada quadrimestre). Ara bé, malgrat que s'ha de cursar en dos quadrimestres, hi ha fins a quatre quadrimestres per fer-ho. Si en quatre quadrimestres no s'ha superat, llavors es pot demanar un any de gràcies, i si l'òrgan responsable del Grau de Matemàtiques ho considera viable s'atorga.

A la base de dades de 494 estudiants (un cop extrets els que no han superat la fase inicial) n'hem suprimit 8 que l'havien superat amb un quadrimestre, perquè segur que corresponien a estudiants amb trasllat d'expedient i força assignatures convalidades. També hi ha 5 dones i 4 homes que han demanat l'any de gràcia i que, per tant, l'han superat amb més de quatre quadrimestres. A la Taula 6 figuren els valors observats d'estudiants dona i estudiants home que han superat la fase inicial en un determinat nombre de quadrimestres.

Tenim que el nombre de quadrimestres que necessita un alumne per a superar la fase inicial segueix una distribució multinomial. Per tant, voldrem veure si els paràmetres de la distribució multinomial dels nois i els de les noies són tots els mateixos o no. Si assumim que Y_1 ens dóna el nombre de quadrimestres emprats per una noia i Y_2 els d'un noi, es té que

$$Y_i \sim Mult(4, p_{i1}, p_{i2}, p_{i3}, p_{i4}), \quad i = 1, 2$$

on $\forall i$ es compleix que $\sum_{j=1}^{4} p_{ij} = 1$.

A partir d'aquí, el test d'hipòtesi a realitzar és:

$$\begin{cases} H_0 : p_{11} = p_{21}, p_{12} = p_{22}, p_{13} = p_{23} \ i \ p_{14} = p_{24} \\ H_1 : \neg H_0 \end{cases} \tag{20}$$

que correspon a un test d'homogeneïtat per veure si els paràmetres de la distribució multinomial de nois i noies són tots iguals o no, és a dir, per veure si segueixen la mateixa distribució ambdues poblacions. Sota la hipòtesi nul·la, el valor esperat de la casella (i,j) es calcula com

$$e_{ij} = N \cdot \hat{p_{i\cdot}} \cdot \hat{p_{\cdot j}}$$

essent N el nombre d'observacions totals de la taula, $\hat{p_{i\cdot}}$ l'estimació de la probabilitat de que una observació pertanyi a la fila i-èssima, que es calcula com el quocient del nombre d'observacions de la fila i dividit pel nombre d'observacions totals, i $\hat{p_{\cdot j}}$ l'estimació de la probabilitat que una observació pertanyi a la columna j-èssima, que s'estima pel quocient d'observacions de la columna j entre el total d'observacions.

Veiem a continuació a la Taula 6 els valors observats i els valors esperats del nombre de quadrimestres que triguen a superar la fase inicial les dones i els homes.

Sexe	**Dones**		**Homes**		**Suma**
Quadrimestres	Observat	Esperat	Observat	Esperat	Observat
2	60	68.48	200	191.52	260
3	24	34.77	108	97.23	132
4	39	22.39	46	62.61	85
5-6	5	2.37	4	6.62	9
Suma	128		358		486

Taula 6: Distribució del nombre de dones i d'homes segons el nombre de quadrimestres emprats a la fase inicial

Realitzem aleshores el test d'homogeneïtat (20). El valor de l'estadístic de prova és

$$X^2 = \sum_{i=1}^{4}\sum_{j=1}^{2} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = 26.647$$

amb 3 graus de llibertat i el resultat del test (20) és que el p-valor és de $6.981 * 10^{-6}$, inferior al nivell de significació 0.05 i, per tant, rebutgem la hipòtesi nul·la de que tots els paràmetres de la distribució multinomial són iguals per nois que per noies.

Per tant, tenim que hi ha diferència entre nois i noies en el nombre de quadrimestres que necessiten per superar la fase inicial. El resultat és que els homes superen la fase inicial amb menys quadrimestres que les dones, com podem observar a la Taula 6.

9. Anàlisi del nombre d'assignatures aprovades a la primera segons el gènere

L'objectiu d'aquesta secció és analitzar si hi ha diferència segons el gènere en el nombre d'assignatures obligatòries del grau que aproven els alumnes a la primera. Això vol dir que no ha fet falta que matriculin l'assignatura una segona vegada, és a dir, han tingut dos intents per aprovar l'assignatura (l'examen final de l'assignatura i l'examen de reavaluació). Tenim que en el Grau de Matemàtiques hi ha un total de 25 assignatures obligatòries i, per poder realitzar l'anàlisi, caldrà eliminar de la base de dades tots aquells alumnes que no hagin superat alguna de les assignatures, quedant un total de 331 alumnes.

Per poder realitzar l'anàlisi, caldrà agrupar el nombre d'assignatures aprovades a la primera en diferents categories. Hem decidit en un principi organitzar-les en 5 categories, on totes les categories tenen la mateixa mida i les podem observar a la Taula 7, així com els valors observats d'assignatures aprovades a la primera pels nois i les noies.

Tenim que el nombre d'assignatures obligatòries aprovades a la primera segueix una distribució multinomial. Si assumim que Y_1 ens dóna la categoria del nombre d'assignatures aprovades a la primera per una noia i Y_2 per un noi, es té que

$$Y_i \sim Mult(5, p_{i1}, p_{i2}, p_{i3}, p_{i4}, p_{i5}), \quad i = 1, 2$$

on $\forall i$ es compleix que $\sum_{j=1}^{5} p_{ij} = 1$.

A partir d'aquí, el test d'hipòtesi a realitzar és:

$$\begin{cases} H_0 : p_{11} = p_{21}, p_{12} = p_{22}, p_{13} = p_{23}, p_{14} = p_{24} \ i \ p_{15} = p_{25} \\ H_1 : \neg H_0 \end{cases} \tag{21}$$

que correspon a un test d'homogeneïtat per veure si els paràmetres de la distribució multinomial de nois i noies són tots iguals o no, és a dir, per veure si segueixen la mateixa distribució ambdues poblacions. Sota la hipòtesi nul·la, el valor esperat de la casella (i,j) es calcula de la mateixa manera que en la Secció 8.

Veiem a continuació a la Taula 7 els valors observats i esperats del nombre d'alumnes a cada categoria d'assignatures aprovades a la primera en funció del gènere.

Sexe	**Dones**		**Homes**		**Suma**
Nombre d'assignatures	Observat	Esperat	Observat	Esperat	Observat
(0,5]	2	1.91	6	6.09	8
(5,10]	2	1.19	3	3.81	5
(10,15]	6	4.3	12	13.7	18
(15,20]	21	11.93	29	38.07	50
(20,25]	48	59.67	202	190.33	250
Suma	79		252		331

Taula 7: Distribució del nombre de dones i d'homes segons el nombre d'assignatures aprovades a la primera

Realitzem aleshores el test d'homogeneïtat. El valor de l'estadístic de prova és

$$X^2 = \sum_{i=1}^{5}\sum_{j=1}^{2} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = 13.654$$

amb 4 graus de llibertat i el resultat del test és que el p-valor és 0.008, inferior al nivell de significació 0.05 i, per tant, rebutgem la hipòtesi nul·la d'aquest test.

Observis que el nombre d'alumnes que hi ha a les diferents categories és molt desigual. Tenim que en les dos primeres categories el valor esperat per les noies és molt petit i, per tant, procedirem a agrupar-les. A més a més, a l'última categoria hi trobem la majoria dels alumnes i per aquest motiu decidim dividir-la en dues noves categories. Això era d'esperar degut a que els estudiants gaudeixen de dos intents per aprovar una assignatura a la primera. Per tant, creem les següents categories (veure Taula 8) per intentar que no hi hagi tanta diferència d'alumnes en els diferents grups. A la Taula 8 observem els valors observats i esperats del nombre d'alumnes per gènere a les noves categories d'assignatures aprovades a la primera.

Sexe	**Dones**		**Homes**		**Suma**
Nombre d'assignatures	Observat	Esperat	Observat	Esperat	Observat
(0,10]	4	3.1	9	9.9	13
(10,15]	6	4.3	12	13.7	18
(15,20]	21	11.93	29	38.07	50
(20,23]	13	10.02	29	31.98	42
(23,25]	35	49.64	173	158.36	208
Suma	79		252		331

Taula 8: Distribució del nombre de dones i d'homes segons el nombre d'assignatures aprovades a la primera

Realitzem novament el test d'homogeneïtat (21). El valor de l'estadístic de prova és

$$X^2 = \sum_{i=1}^{5}\sum_{j=1}^{2} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = 17.11$$

amb 4 graus de llibertat i el resultat del test és que el p-valor és 0.002, inferior al nivell de significació 0.05 i, per tant, rebutgem la hipòtesi nul·la.

Per tant, tenim que hi ha diferència entre nois i noies en el nombre d'assignatures obligatòries que aproven a la primera en el Grau de Matemàtiques. Podem observar a la Taula 8 que un nombre més elevat d'homes aprova 24 o 25 assignatures a la primera respecte al valor esperat, mentre que de dones hi ha menys de les esperades.

10. Anàlisi de la probabilitat de no abandonar el grau

L'objectiu d'aquesta secció és analitzar la probabilitat que tenen els estudiants de no abandonar el Grau de Matemàtiques tenint en compte la informació de la qual disposem una vegada accedeixen a la facultat. Per fer-ho, ha sigut necessari crear la variable *no abandonar* explicada a la Secció 4, la qual pren valors 1 pels estudiants que no han abandonat el grau i 0 per aquells que si que ho han fet. La base de dades per realitzar l'anàlisi consta de 585 estudiants que han accedit al Grau de Matemàtiques entre els anys 2009 i 2017. A continuació, procedim a realitzar l'anàlisi descriptiva de les dades, d'on obtenim la Taula 9 que mostra el nombre d'alumnes que no han abandonat el grau i els que sí ho han fet segons el gènere i si els alumnes són *CFIS* o no ho són. D'aquesta taula es desprèn que en general es estudiants *CFIS* no abandonen, mentre que dels no *CFIS* abandonen una mica més d'un 40%.

Sexe	**Dones**		**Homes**		**Suma**		
	No aban.	Aban.	No aban.	Aban.	No aban.	Aban.	Total
No CFIS	74	67	137	101	211	168	379
CFIS	24	4	166	12	190	16	206
Suma	98	71	303	113	401	184	585
	169		416				

Taula 9: Distribució d'alumnes que han abandonat o no segons el gènere i si són CFIS o no

A continuació, volem realitzar un model que predigui la probabilitat de no abandonar d'un estudiant quan entra al Grau de Matemàtiques. Sabem que la probabilitat de no abandonar segueix una distribució Binomial amb paràmetres p_i, que dependrà de les característiques de l'estudiant, i n=1 (en particular, tenim una distribució Bernoulli). Per tant, realitzarem un model lineal generalitzat amb resposta Binomial i utilitzarem les funcions d'enllaç més freqüents d'aquesta distribució, que són *logit*, que es correspon a la funció d'enllaç canònica pels models lineals generalitzats amb resposta Binomial, *probit* i *c-log-log* per veure amb quina obtenim millors resultats. Com que els resultats dels models obtinguts amb les diferents funcions d'enllaç són molt semblants, ens hem quedat amb la funció *logit*, ja que simplifica la interpretació del model i fa els càlculs més senzills.

Realitzem un primer model que utilitzi com a variables explicatives el gènere, si l'alumne és *CFIS* o no ho és, la diferència entre la nota d'accés de l'alumne i la nota d'accés mínima del curs i *l'any d'entrada*. El resultat és que la variable gènere no és significativa i també trobem que només l'any 2013 és significativament diferent de l'any que es pren com a referència, que és el 2009. A més a més, tenim que la deviança utilitzant el model nul és de 728.53, mentre que la deviança amb el nostre model és de 598.21. També és important comentar que el AIC pren el valor de 622.21, ja que serà important per comparar models que utilitzin un nombre diferent de paràmetres, ja que en el seu càlcul sí que es té en compte el nombre de paràmetres a diferència del que succeeix amb la deviança.

Observem aleshores que el AIC del mateix model sense utilitzar *l'any d'entrada* és de 632.43, és a dir, no és gaire més alt. Com que el model no empitjora gaire, només hi ha un any diferent respecte l'any de referència i per poder predir la probabilitat de que un estudiant no abandoni amb la informació que tenim quan aquest accedeix al Grau de Matemàtiques necessitem no tenir en compte l'any d'entrada, optem aleshores per eliminar la variable *any d'entrada*.

Realitzem doncs un nou model que utilitza les variables explicatives del gènere, si l'estudiant és *CFIS* o no i la *diferència de notes*. El resultat d'aquest model és que el gènere no és una variable significativa i, per tant, l'eliminem. Per tant, tenim un model que utilitza com a variables explicatives si l'alumne és *CFIS* o no i la diferència de notes. Podem observar a la Figura 20 que la deviança residual del model és 626.11 i el AIC és de 632.11, que són semblants al primer model que havíem realitzat i aquest nou model és més senzill i utilitza molts menys paràmetres que l'anterior. Calculem aleshores l'estimació del paràmetre de dispersió

$$\hat{\phi} = \frac{\sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i(1-\hat{\mu}_i)}}{n-p} = 1.027414$$

on observem que és pràcticament 1 i, per tant, concloem que té sentit utilitzar la distribució Binomial per a la variable resposta.

```
Call:
glm(formula = abandonat ~ CFIS + diferencia_notes, family = binomial(link = "logit"),
    data = dades)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4327  -1.1558   0.4118   1.0000   1.2509

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -0.17110    0.18399  -0.930  0.35238
CFISSI           2.15146    0.28220   7.624 2.46e-14 ***
diferencia_notes  0.15578    0.05978   2.606  0.00916 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 728.53  on 584  degrees of freedom
Residual deviance: 626.11  on 582  degrees of freedom
AIC: 632.11

Number of Fisher Scoring iterations: 5
```

Figura 20: Resum del model (22)

Finalment, provem d'afegir la variable interacció *CFIS*diferència de notes*, però el resultat és que aquesta nova variable no és significativa.

Per tant, el model final que hem obtingut és el següent:

$$log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 \delta_{CFIS} + \beta_2 \text{Diferència de notes} \tag{22}$$

En particular, tenim que amb els coeficients observats a la Figura 20 el model pels estudiants *CFIS* és

$$\text{Estudiant CFIS:} \quad log\left(\frac{p_i}{1-p_i}\right) = 1.98 + 0.16 \cdot \text{Diferència de notes}$$

i el model pels estudiants no *CFIS* és

$$\text{Estudiant no CFIS:} \quad \log\left(\frac{p_i}{1-p_i}\right) = -0.17 + 0.16 \cdot \text{Diferència de notes}$$

Per veure que podem considerar aquest model com un bon model, adjuntem la gràfica dels residus de Pearson contra el valor del predictor lineal.

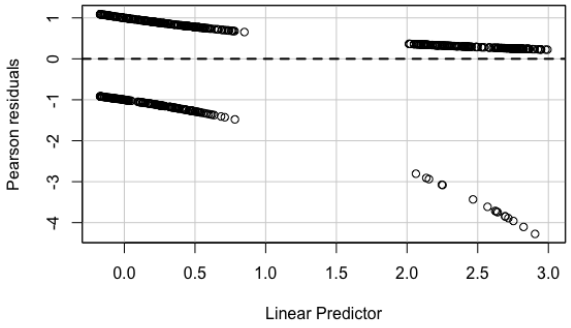


Figura 21: Gràfica de residus del model (22)

En la Figura 21 podem observar clarament que hi ha dues poblacions separades, les quals es corresponen als estudiants que són *CFIS* (els punts de la dreta) i als que no ho són (els punts de l'esquerra). A més a més, observem que els residus de Pearson són prou petits excepte pels casos que es troben a baix a la dreta, que es corresponen als 16 estudiants *CFIS* que tenim a les dades que han abandonat, ja que el model no preveu que els estudiants *CFIS* abandonin degut a l'alt valor del coeficient d'aquesta variable. Per tant, pel que fa als residus, podem considerar el nostre model com un bon model.

Finalment, anem a veure quins són els valors predits de *no abandonar* dels estudiants utilitzant una *diferència de notes* mitja, és a dir, fent la mitjana de les diferències de notes que tenim dels estudiants *CFIS* i dels que no ho són. Tenim que la *diferència de notes* mitjana pels estudiants *CFIS* és de 3.39 i, per tant, la probabilitat de no abandonar d'un estudiant *CFIS* amb aquesta diferència de notes és

$$\text{Estudiant CFIS:} \quad \hat{p} = \frac{e^{1.98+0.16\cdot 3.39}}{1+e^{1.98+0.16\cdot 3.39}} = 0.92$$

En canvi, pels estudiants que no són *CFIS* tenim que la *diferència de notes* mitja és de 2.58 i, per tant, la probabilitat de no abandonar d'aquests estudiants tenint aquesta *diferència de notes* és

$$\text{Estudiant no CFIS:} \quad \hat{p} = \frac{e^{-0.17+0.16\cdot 2.58}}{1+e^{-0.17+0.16\cdot 2.58}} = 0.56$$

Per tant, podem concloure que la probabilitat de no abandonar el Grau de Matemàtiques varia molt segons si els estudiants són *CFIS* o no ho són, sent molt difícil que un estudiant *CFIS* abandoni el Grau i, en canvi, baixant del 50% de probabilitat de no abandonar dels estudiants que no ho són segons quina sigui la seva *diferència de notes*.

11. Conclusions

L'objectiu del treball consistia en estudiar els resultats acadèmics dels estudiants del Grau de Matemàtiques per veure si el gènere hi té influència i, en cas afirmatiu, de quina manera ho fa. Amb aquest objectiu, després de realitzar les anàlisis hem arribat a les següents conclusions:

- Hem pogut observar que no hi ha una diferència significativa en la nota amb la que accedeixen a la facultat els nois i les noies. Això vol dir que, pel que fa al nivell de Batxillerat tan com de Selectivitat, els nois i les noies que accedeixen al Grau de Matemàtiques el tenen semblant.
- Hem vist que hi ha diferència en els resultats dels nois i de les noies a la *prova de nivell* que es realitza als estudiants que accedeixen al grau abans de començar les classes. Tenim que la mitjana de la nota de la *prova de nivell* dels homes és gairebé 1 punt més alta que la de les dones pels estudiants que han realitzat la prova entre els anys 2009 i 2019.
- Quan intentem crear un model que explica la nota *mitjana de la fase inicial* dels alumnes a partir de la seva *nota d'accés*, la *nota d'accés mínima d'aquell curs*, el gènere i si l'alumne és *CFIS* o no ho és, el resultat és que els homes treuen 0.4 punts més que les dones. Tot i això, afecta molt més a la nota *mitjana de la fase inicial* el fet de ser *CFIS* o no que el gènere de l'alumne, ja que aquesta variable augmenta la predicció en 1.2 punts.
- Si intentem crear un model que explica la nota *mitjana de la fase inicial* a partir de la diferència entre la *nota d'accés* i la *nota d'accés mínima del curs*, la nota de la *prova de nivell*, el gènere i si l'alumne és *CFIS* o no ho és, el resultat és que la variable gènere no afecta al model. Això, però, no vol dir que no hi hagi diferències entre nois i noies a la nota *mitjana de la fase inicial*, ja que la influència del gènere podria ser que es tingués en compte en les diferències de la nota de la *prova de nivell*.
- Hem comprovat que la distribució multinomial del *nombre de quadrimestres necessaris per superar la fase inicial* de les dones és diferent que la que segueixen els homes.
- Hem comprovat que homes i dones no segueixen la mateixa distribució multinomial pel que fa al *nombre d'assignatures obligatòries aprovades a la primera*. Hem vist que la probabilitat d'aprovar un número molt elevat d'assignatures a la primera és més alta pels homes que per les dones.
- Per acabar, podem concloure que no hi ha diferència entre els nois i les noies pel que fa a la probabilitat de *no abandonar* el Grau, així que, per molt que potser les noies no aprovin tantes assignatures a la primera, se n'acaben sortint i finalitzen el grau amb la mateixa probabilitat que els nois. En canvi, hem observat que en aquesta variable si que hi ha molta diferència en si l'estudiant és *CFIS* o no ho és, sent la probabilitat de no abandonar dels primers vora del 90%, mentre que dels estudiants que no són *CFIS*, depenent de la seva *nota d'accés*, pot ser més probable que abandonin que que no ho facin.

Finalment, voldríem dir que hi ha moltes altres variables que poden afectar les variables resposta estudiades en aquest TFG, com per exemple les *hores que un estudiant dedica a l'estudi d'una assignatura*, el *professor* que imparteix l'assignatura, etc. Tenir en compte aquestes variables hauria requerit d'una elevada planificació de la recollida de dades i de fer el seguiment dels estudiants al llarg d'uns quants anys. No obstant, l'anàlisi de les dades recopilades de forma automàtica a la FME al llarg dels anys considerats ens ha permès tenir uns idea prou bona del comportament dels estudiants amb perspectiva de gènere.

Referències

[1] Annette J. Dobson. *An introduction to Generalized Linear Models*, 1st ed. Chapman & Hall, 1990.

[2] P. McCullagh and J.A. Nelder. *Generalized Linear Models*, 2nd ed. Chapman & Hall, 1989.

[3] Marta Pérez-Casany. *Apunts de l'assignatura d'Estadística del Grau de Matemàtiques de la FME*, 2018.

[4] Marta Pérez-Casany. *Apunts de l'assignatura Models Lineals i Lineals Generalitzats del màster MESIO UPC-UB*, 2019.

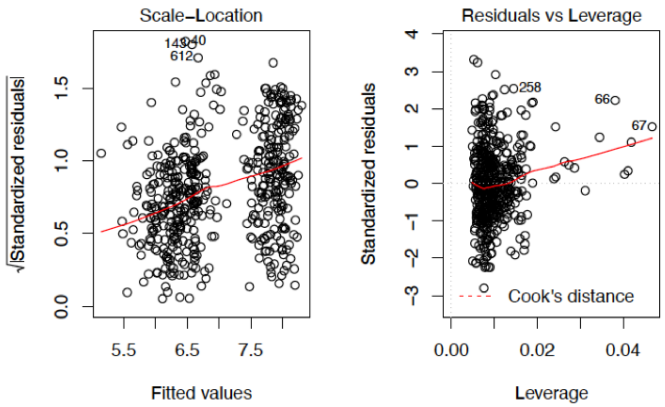
[5] Ugarte, M. D., Militino, A. F. and Arnholt, A. T. *Probability and Statistics with R*, 2nd ed. Chapman & Hall, 2015.

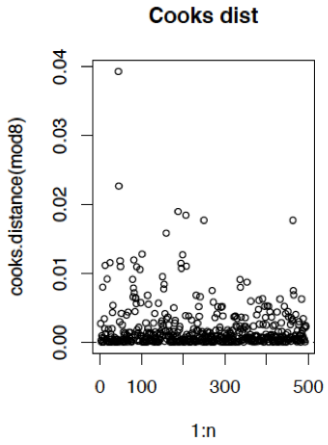
A. Codi de les anàlisis

A.1 Secció 6

A continuació observem el codi utilitzat per realitzar el model final de la Secció 6, així com les gràfiques que no hem inclòs en el cos del treball.

```
mod8 <- lm(mitjana_FI ~ sexe + nota_acces_14 + nota_acces_min_curs_14 + CFIS, data = dades)
summary(mod8)
vif(mod8)
Anova(mod8)
par(mfrow=c(1,2))
plot(mod8)
plot(1:n,cooks.distance(mod8),cex=.75,main="Cooks dist")
abline(h=c(0),lty=2)
predict(mod8,newdata=data.frame(sexe=c("HOME","DONA"),nota_acces_14=c(13.13,13.13),
nota_acces_min_curs_14=c(10.548,10.548),CFIS=c("NO","NO")))
predict(mod8,newdata=data.frame(sexe=c("HOME","DONA"),nota_acces_14=c(13.13,13.13),
nota_acces_min_curs_14=c(10.548,10.548),CFIS=c("SI","SI")))
```

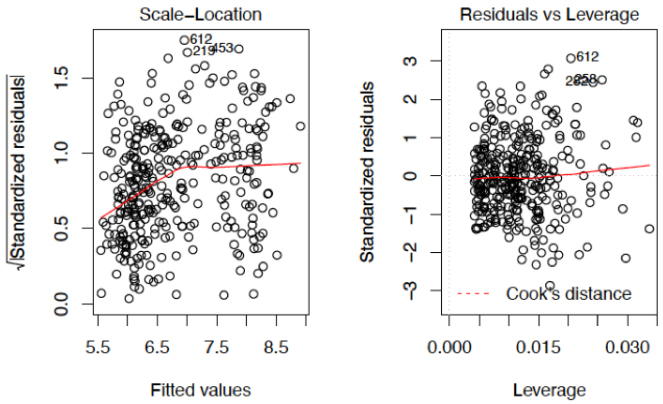


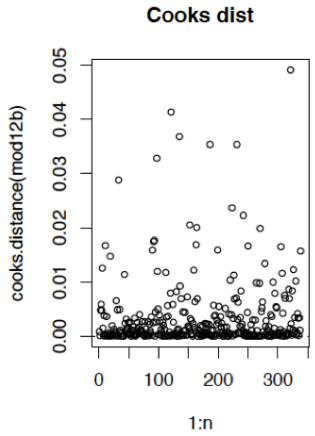


A.2 Secció 7

A continuació observem el codi utilitzat per realitzar el model final de la Secció 7, així com les gràfiques que no hem inclòs en el cos del treball.

```
mod12b <- lm(mitjana_FI ~ provaNivell2 + diferencia_notes + CFIS, data = dades)
summary(mod12b)
Anova(mod12b)
plot(mod12b)
vif(mod12b)
plot(1:n,cooks.distance(mod12b),cex=.75,main="Cooks dist")
abline(h=c(0),lty=2)
predict(mod12b,newdata=data.frame(provaNivell2=c(52.5625,52.5625),diferencia_notes=c(2.63,2.63),
CFIS=c("SI","NO")))
```





A.3 Secció 8

A continuació observem el codi utilitzat per realitzar la Secció 8, així com la taula de probabilitats marginals del nombre de quadrimestres necessaris per superar la fase inicial segons el gènere.

```
dades$nombre_quad_FI <- cut(dades$nombre_quad_FI,c(1,2,3,4,6))
taula <- table(dades$nombre_quad_FI, dades$sexe)
addmargins(taula)
prop.table(taula)
prop.table(taula,2)
x=chisq.test(taula)
x$expected
x$observed
```

```
             DONA       HOME
(1,2] 0.46875000 0.55865922
(2,3] 0.18750000 0.30167598
(3,4] 0.30468750 0.12849162
(4,6] 0.03906250 0.01117318
```

A.4 Secció 9

A continuació observem el codi utilitzat per realitzar la Taula 8 de la Secció 9, així com la taula de probabilitats marginals del nombre d'assignatures aprovades a la primera segons el gènere.

```
dades$categories2_assigs_aprovades_1 <- cut(dades$nombre_assigs_aprovades_1,
c(0,10,15,20,23,25))
dades$categories2_assigs_aprovades_1 <-
factor(dades$categories2_assigs_aprovades_1)
taula2 <- table(dades$categories2_assigs_aprovades_1, dades$sexe)
addmargins(taula2)
prop.table(taula2)
prop.table(taula2,2)
x2=chisq.test(taula2)
x2$expected
x2$observed
```

```
              DONA       HOME
(0,10]  0.05063291 0.03571429
(10,15] 0.07594937 0.04761905
(15,20] 0.26582278 0.11507937
(20,23] 0.16455696 0.11507937
(23,25] 0.44303797 0.68650794
```

A.5 Secció 10

A continuació observem el codi utilitzat per realitzar el model final de la Secció 10.

```
mod6 <- glm(abandonat~ CFIS + diferencia_notes,family=binomial(link="logit"),
data=dades)
summary(mod6)
# Calculem l'estimació de phi
PS<-sum(residuals(mod6,type="pearson")^2)
PS/mod6$df.res
residualPlot(mod6,smooth=F)
predict(mod6,newdata=data.frame(CFIS=c("NO","SI"),diferencia_notes=c(2.58,3.39)),
type="response")
```